Department of Computer Science
Series of Publications A
Report A-2020-5

Epistemological Approach to Dependability of Intelligent Distributed Systems

Heimo Laamanen

Doctoral dissertation, to be presented for public examination with the permission of the Faculty of Science of the University of Helsinki, in Auditorium 2, Metsätalo, Helsinki, Finland, on the 26th of June, 2020 at 12 o'clock.

University of Helsinki
Finland

Supervisors
Jussi Kangasharju, University of Helsinki, Finland
Markus Lammenranta, University of Helsinki, Finland

Pre-examiners
Ahti–Veikko Pietarinen, Nazarbayev University, Kazakhstan
Petri Ylikoski, University of Helsinki, Finland

Opponent
Erkki Sutinen, University of Turku, Namibia

Custos
Jussi Kangasharju, University of Helsinki, Finland

Contact information

Department of Computer Science
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014 University of Helsinki
Finland

Email address: info@cs.helsinki.fi
URL: http://cs.helsinki.fi/
Telephone: +358 2941 911

ISSN 1238-8645
ISBN 978-951-51-6201-4 (paperback)
ISBN 978-951-51-6202-1 (PDF)
Helsinki 2020
Unigrafia

Epistemological Approach to Dependability of Intelligent Distributed Systems

Heimo Laamanen

Department of Computer Science
P.O. Box 68, FI-00014 University of Helsinki, Finland
Heimo.Laamanen@helsinki.fi

PhD Thesis, Series of Publications A, Report A-2020-5
Helsinki, June 2020, 204 + 113 pages
ISSN 1238-8645
ISBN 978-951-51-6201-4 (paperback)
ISBN 978-951-51-6202-1 (PDF)

Abstract

Recent and expected future developments in the domains of artificial intelligence, intelligent software agents, and robotics will create a new kind of environment where artificial entities and human beings seamlessly operate together to offer services. The users of these services may not necessary know whether the service is actually offered by a human being or an artificial entity. This kind of environment raises a requirement for using a joint terminology between human beings and artificial entities, especially in the domain of the epistemic quality of information. The epistemic quality of information will play an important role in this kind of intelligent distributed systems. One of the main reasons is that it affects the dependability of those systems.

Epistemology is the study of knowledge and justified belief including their nature, sources, limits, and forms. Human beings have been interested in epistemology since the times of ancient Greece, as knowledge is seen to be an important factor of human beings' actions and success in the actions. We are of the opinion that the scene of epistemology is changing more than ever before: artificial intelligence has entered into the domain. In this thesis we argue that first, an intelligent software entity is capable of having beliefs and second, both knowledge and justified belief will be important factors in the dependability of AI–based agents' actions and success in the actions.

We carry out a theoretical analysis of the epistemological concepts—belief, justified belief, and knowledge—for the context of intelligent software agents and dependable intelligent distributed systems. We introduce enhanced definitions of justified belief and knowledge, which we call Pragmatic Process Reliabilism. These definitions can be adopted into dependable intelligent distributed systems.

We enhance the dependability taxonomy in order to cope better with the situations created by learning and the variation of the epistemic quality of information. The enhancements comprise the following concepts: attributes (skillfulness, truthfulness, and serveability), fault classes (training fault and learning fault), failure (action failure and observed failure), and means (relearning and retraining).

We develop a theoretical framework (Belief Description Framework – BDF) to perceive, process, and distribute information in order to verify that our ideas can be implemented. We model the framework using Unified Modelling Language in order to demonstrate its applicability for implementation. First, we define relationships between epistemological concepts and software entities (classes). Second, we show that information, belief, justified belief, and knowledge can be specified as classes and instantiated as objects. The Information class defines the environment—a kind of information ecosystem—of information. It is the central point. It has relationships with other classes: Proposition, Presentation, EpistemicQuality, Warrant, Security, Context, and ActorOnInformation. Third, we specify some important requirements for BDF. Fourth, we show by modelling BDF using the UML modelling method that BDF can be specified and implemented.

Computing Reviews (2012) Categories and Subject Descriptors:

Computer system organization
→ Dependable and fault–tolerant systems and networks

Computing methodologies
→ Artificial intelligence
→ Philosophical/theoretical foundations of artificial intelligence

Information systems
→ Information retrieval
→ Document representation → Content analysis and feature selection
→ Evaluation of retrieval results → Relevance assessment

General Terms:
Epistemology, Dependability, Distributed Systems, Software Agents, Belief, Justified Belief, Knowledge

Additional Key Words and Phrases:
Justification Theory, Knowledge Theory, Dependability Taxonomy

Acknowledgements

As my studying history is quite long, I want to express my gratitude to many persons. First of all, I am very grateful to my supervisors Jussi Kangasharju and Markus Lammenranta for their excellent guidance in the domains of computer science and philosophy respectively. Jussi Kangasharju guided me throughout my PhD research, especially in the domains of intelligent distributed systems and Belief Description Framework. Markus Lammenranta played an important role in guiding me through the difficulties of epistemology and presented many ideas that improved this thesis a lot. I owe also many thanks to Raul Hakli, Markku Kojo and the late Kimmo Raatikainen. I express my deepest gratitude to my long time mentor Timo Alanko. Without his inspiring and supportive mentoring during all my very many studying years this thesis would never have been written.

I thank the pre–examiners of this thesis, Ahti–Veikko Pietarinen and Petri Ylikoski for their helpful and valuable comments.

I owe many thanks to the Department of Computer Science for providing an excellent environment for studies and research. For example, Marina Kurtén helped me to improve the language in this thesis, and Pirjo Moen guided and assisted me through the administrative tasks in addition to the improvement of the layout of this thesis. And once more I praise Jussi Kangasharju, who accepted my kind of old–timer, 'never–ending student' as his PhD student.

I would like to thank Ari Kinnunen for his ideas of the health care scenario.

Lastly but not least, this thesis would not have been possible without the support of my friends and especially my wife Sirkka Laamanen.

Espoo, June 2020
Heimo Laamanen

Conventions

In this thesis we use several terms that have a different meaning in the disciplines of computer science and philosophy. Therefore, we use the following conventions to make the distinction between the meanings of computer science and philosophy:

1. Superscript "p" is used when a term has the philosophical meaning. For example, reliability^p refers to the philosophical meaning of the term *reliability*.[1]

2. Superscript "c" is used when a term has the meaning specified in computer science. For example, reliability^c refers to the meaning of the term *reliability* in computer science.

3. Subscript "bdi" is used when we refer to the belief–desire–intention type of intelligent software agent. For example, ISA$_{bdi}$ is one type of intelligent software agent.

Use of fonts:

1. *Italics* is used to emphasize an important term, a key part of a text, or other points worth of special attention.

2. **Bold** is used to emphasize a term or an abbreviation.

3. Small capitals is used in definitions defined by us.

4. *Slanted shape* is used in quotations and in the names of classes and objects.

[1] See Appendix Terminology.

Abbreviations:

1. ADC — Agent Driven Car
2. AI — Artificial Intelligence
3. BDF — Belief Description Framework
4. BDI — Belief, Desire, and Intention
5. CTL — Computation Tree Logic
6. CTM — Computational Theory of Mind
7. DNS — Domain Name System
8. DIDS — Dependable Intelligent Distributed System
9. FIPA — Foundation for Intelligent Physical Agent
10. GOFAI — Good Old Fashioned AI
11. DAML — DARPA Agent Markup Language
12. DL — Direct Semantics
13. EMA — Emergency Medical Assistance
14. HDC — Human Driven Car
15. IDS — Intelligent Distributed System
16. ISA — Intelligent Software Agent
17. ISA_{bdi} — Intelligent Software Agent based on BDI architecture
18. JTB — Justified True Belief
19. OWL — Web Ontology Language
20. PPR — Pragmatic Process Reliabilism
21. RDF — Resource Description Framework
22. THIS — Travellers' Health and Insurance Service
23. TIS — Traffic Information Service
24. TMA — Travellers' Medical Assistance

25. UDDI — Universal Description Discovery and Integration
26. UML — Unified Modeling Language
27. URL — Universal Resource Locator
28. VM — Virtual Machine
29. VMF — Virtual Machine Functionalism
30. WSDL — Web Services Description Language
31. XML — Extensible Markup Language.

I have used the following tools to write this thesis:

1. Texmaker/TeXstudio: the latex editors to write the text and create the PDF copy of this thesis.
2. JabRef/kBibTex: The bibliography reference managers to manage the reference database and the references.
3. StarUML: The software modeller to develop the UML models.
4. LibreOffice Draw: The tool to draw figures.
5. Protege: The ontology editor to develop the ontologies.
6. Calibre: The E–book manager to manage the collection of articles and books that are referred to in this thesis.

License information of figures:

1. Robot holding a book:
   http://creativecommons.org/licenses/publicdomain/ by Ikebanto.
2. Head and brain: Creative Commons 4.0 BY–NC.

A note on URLs in bibliography references:

We have checked the correctness of URLs. However, some URLs may change or disappear as time passes. In that case, please, utilize Internet search services to locate the referred article.

If you begin with Computer Science,

you will end with Philosophy.

William J. Rapaport

Contents

List of Figures **xvii**

List of Tables **xix**

1 Introduction **1**
1.1 Motivation and Problem Statement 4
1.2 Contributions 6
1.3 Structure of Thesis 7

2 Background and Overview **9**
2.1 Introduction to Dependability Issues 9
2.1.1 Scenarios 9
Faked White House Bomb Tweet Causes Stock Market Panic 10
A Tourist having an Accident in a Foreign Country . 12
Traffic Information Service 14
2.1.2 Dependability Theory 17
Basic Concepts and Taxonomy 17
2.2 Intelligent Distributed Systems 21
2.2.1 Artificial Intelligence 23
GOFAI 23
Connectionism 24
Hybrid Approach 25
Intelligent Software Agents 26
Representation of Semantic Information 30
2.2.2 Knowledge and Justified Belief in Dependable Intelligent Distributed Systems 32
2.2.3 Logical Issues of Knowledge, Justified Belief, and Belief 37
Epistemic Logic 40
Epistemic Logic of Single Agent 41

Epistemic Logic of Multiple Agents 43
Epistemic Logic of Justification 44
Summary of Logical Issues 47

3 Six Concepts **49**
3.1 Introduction to the Six Concepts 49
3.2 Epistemic Value 56
3.3 Truth 61
3.3.1 Truth Theories 65
The Coherence Theory of Truth 65
The Pragmatic Theory of Truth 66
The Redundancy Theory of Truth 66
The Correspondence Theory of Truth 67
The Identity Theory of Truth 68
3.3.2 Speech Act Theory and ISA Asserting Propositions . 68
3.3.3 Thoughts about Truth and ISA_{bdi} 70
3.3.4 Conclusions about Truth in the Context of ISA_{bdi} . 73
3.4 Belief 73
3.5 Justified Belief 78
3.5.1 Internalism and Externalism 80
3.5.2 Foundationalism about Justified Belief 82
3.5.3 Coherentism about Justified Belief 84
3.5.4 Evidentialism about Justified Belief 86
3.5.5 Reliabilism about Justified Belief 87
3.5.6 Testimony about Justified Belief 91
3.5.7 Conclusion about Justified Belief in the context of ISA_{bdi} 98
3.6 Knowledge 99
3.6.1 Testimony about Knowledge 104
3.6.2 Causal Theory about Knowledge 107
3.6.3 Virtue Epistemology about Knowledge 107
3.6.4 Knowledge First about Knowledge 108
3.6.5 Reliabilism about Knowledge 109
3.6.6 Conclusion about Knowledge in the context of ISA_{bdi} 115
3.7 Trust 116
3.8 Possible Objections 121
Objection 1: Anthropomorphism 121
Objection 2: Joint Epistemic Theories 122
Objection 3: Pragmatic Process Reliabilism as Joint Epistemic Theory 123
Objection 4: Implementability 124

3.9 Summary of Six Concepts 124
Truth 124
Trust and Trustworthiness 124
Summary of Definitions 125
Conclusions of Six Concepts 126

4 Belief as Dependability Factor 129
4.1 Justifiably be Trusted 129
4.2 Evaluation of Epistemic Quality of Belief 132
4.2.1 Sources of Beliefs 132
4.2.2 Evaluation of Consequences 135
4.3 Summary of Belief as Dependability Factor 137

5 Enhancement to Dependability Taxonomy 139
5.1 Issues of Dependability Taxonomy 139
5.2 Attributes 141
5.3 Faults 143
5.4 Failures 145
5.5 Means 146
5.6 Discussion about New Attributes 146
5.7 Problems of Implementing Dependability Concerning Epistemic Quality of Information 147
5.8 Summary of Dependability Taxonomy 147

6 Belief Description Framework 153
6.1 Associations between Epistemic Quality and Software Entities 154
6.2 Requirements for BDF 166
Information 169
Information Source 169
Information Processing 170
Information Warrant 172
Possible Worlds 173
6.3 Specifications of BDF 174
6.3.1 Classes and Objects 174
6.3.2 Collaboration 177
6.4 BDF and DIDS 181
6.5 Summary of BDF 181

7 Conclusions 187

References 193

Appendices **204**
Terminology 207
Belief Description Framework 215
Discussions on Evaluating Epistemic Quality of Beliefs 247
Is It Time to Get Out of the Chinese Room? 298

List of Figures

2.1 A scenario of traffic information service. 15
2.2 UML use case of traffic information service. 15
2.3 An example of TIS utilizing a certification service. 16
2.4 Dependability taxonomy. 19
2.5 An example of an intelligent distributed system. 22
2.6 BDI architecture. 29
2.7 Different contexts of propositions. 32

3.1 Human belief and ISA belief. 52
3.2 Classification of information. 55
3.3 Justification. 95
3.4 Truth condition. 102
3.5 Cases of trust. 118

4.1 Justifiably be trusted. 130
4.2 High level scheme of context of epistemic evaluation. 131
4.3 Sources of information of ISA. 133
4.4 Evaluation of consequences. 137

6.1 Classes of the epistemic quality of information. 154
6.2 Associations of information. 155
6.3 Information concepts instantiated as stereotype classes of virtual machine functionality. 158
6.4 Information class. 159
6.5 An example of a belief object. 160
6.6 An example of a justified belief object. 163
6.7 An example of a knowledge object. 165
6.8 Information object structure of knowledge. 167
6.9 An example of use case. 168
6.10 Use case: perceive information. 170
6.11 Use case: evaluation of information. 171

6.12 BDF classes. 175
6.13 BDF epistemic quality class. 176
6.14 BDF instance of information class: justified belief. 178
6.15 BDF sequence diagram of evaluation of information. 180
6.16 BDF activity diagram of evaluation of information – apriori. 182
6.17 BDF activity diagram of evaluation of information – warrant. 183
6.18 BDF activity diagram of distributing information. 184

List of Tables

4.1 Summary of sources of belief. 135
4.2 Relevant possible worlds of Traffic Information Service and evaluated reliability^p requirements for declarations. 136

5.1 Fault classes. 144
5.2 Summary of improvements on dependability taxonomy. . . 151

Chapter 1

Introduction

In the future more and more information services are provided by cooperative groups of human beings and intelligent software agents (hereinafter ISA) based on artificial intelligence (hereinafter AI) [2]. The users of these services may not necessary know, or do not even want to know, whether a service is actually offered by a human being or an artificial entity. When a user of information—either a human being, a robot, or an ISA—obtains a piece of information in order to utilize it, then the following questions can be raised: What is the epistemic quality of the piece of information? Is it knowledge^p, justified belief^p, belief^p, nonsense, or what? Should the user rely on it when planning and carrying out further actions? We argue that epistemology provides proper methods to answer these questions also in the domains of AI and computer science. And the joint context of human beings and ISAs also raises a requirement for using the same terminology between human beings and artificial entities, especially in the domain of the epistemic quality of information.

Epistemology is the study of knowledge^p and justified belief^p including their nature, sources, limits, and forms. Human beings have been interested in epistemology since the times of ancient Greece, as knowledge^p is seen to be an important factor of human beings' actions and success in the actions. Now, the scene of epistemology is changing more than ever before: AI has entered into the domain. In this thesis we argue that knowledge^p and justified belief^p will also be important factors of AI–based agents' actions and success in their actions.

The epistemic quality of information is related to the dependability theory of computer science because it affects the dependability of intelligent distributed systems (hereinafter IDS). Incorrect, false input information most probably causes a failure of a service provided by IDS. There are two major aspects that we need to analyse and synthesize in order to establish

a firm foundation of the epistemic quality of information for the dependability theory. The aspects deal with the existing dependability theory of computer science and the concepts of information, belief^p, justified belief^p, knowledge^p, truth, and trustworthiness in epistemology.

The main research questions in the domain of computer science are as follows:

1. Is it possible to design and implement an ISA which complies with human beings' epistemic concepts of information?

2. In which cases does an artificial epistemic agent deal with knowledge^p, justified belief^p, and belief^p when it perceives or distribute information in the context of IDS?

3. What is the relationship between trust and the epistemic quality of information in the contexts of ISA and IDS?

4. What are the grounds for an artificial epistemic agent to trust information provided by IDS?

5. What kind of enhancements are required to the dependability taxonomy of computer science so that it better address the issues related to learning and the varying epistemic quality of information?

The main research questions in the domain of epistemology are as follows:

1. Is it possible for an artificial entity, such as ISA, to have beliefs^p, justified beliefs^p, and knowledge^p?[1]

2. What kind of concepts are knowledge^p, justified belief^p, belief^p, truth, and trustworthiness^p in the contexts of ISA and IDS?

3. Is it possible and beneficial to define joint definitions of belief^p, justified belief^p, and knowledge^p for both artificial entities and human beings?

In this thesis we have mainly a theoretical approach to the above questions, because practical implementations and the proofs of developed concepts would require a multidisciplinary (artificial intelligence, human computer interaction, epistemology, psychology, and sociology) project.[2]

[1] This is related to the issue of anthropomorphism.

[2] This kind of project requires a lot of manpower which is outside the possibilities of this research project. The implementation and the proofs of concepts will be the topic of the future research.

Recent developments in AI, ISAs, and robotics have shown that artificial entities do exhibit human–like behaviour and therefore indicating a possibility to have beliefs^p, justified beliefs^p, and knowledge^p. In addition, foundational questions and challenges in the development of AI are philosophical in nature dealing with concepts of knowledge^p, representation, and action. In the year 1980 John R. Searle raised a long standing and severe dispute about the capabilities of computer systems to be a mind; thus, to understand, to have intentions^p, to have beliefs^p, etc. In his article *Minds, Brain, and Program* he used the now famous Chinese Room argument to state the following main theses: (1) Intentionality in human beings is created by causal features of the brain and (2) instantiating a computer program is never by itself a sufficient condition of intentionality [128]. One of the main objectives was the view that formal computations on symbols could not produce thought. The reason is that there is no way to attach any meaning to the formal symbols because syntax and internal connections are insufficient for semantics [27]. We argue that *artificial entities such as ISAs* are capable to have, for example, beliefs^p, justified beliefs^p, and knowledge^p. We will discuss our arguments about these issues in more detail in Chapter 3.

Our intention in this thesis is to establish a solid, theoretical foundation for belief^p, justified belief^p, and knowledge^p for the context of IDS, where an ISA provides—possibly in co–operation with human beings—human beings and other ISAs with dependable information when acting on behalf of human beings in dependable intelligent distributed systems (hereinafter DIDS). This will comprise a requirement analysis, natural language (as a meta–language) descriptions of justification theories, truth theories, and knowledge theories.

We discuss the epistemological concepts of belief^p, justified belief^p, and knowledge^p, so that they can be better understood in the contexts of ISA and IDS. We also enhance the concepts of justified belief^p and knowledge^p and adapts these concepts for the contexts of ISA and DIDS. The adaptation of the above–mentioned epistemic theories means to select, modify, or define the theories to be proper in the context of ISA; hence, to be applicable for the theories of dependable computing. The adaptation introduces a new viewpoint to epistemology: traditional epistemology is the study of concepts used by human beings, but our approach is also to study how to implement those existing epistemological concepts in the context of artificial entities. In addition, the concepts of information, truth, and trustworthiness are explored in order to form a firm ground to discuss the epistemological concepts.

We utilize in this thesis the concept of ISA as the abstract model of an intelligent software entity and especially the version of ISA where the concept is based on a *Belief–Desire–Intention* (hereinafter BDI) architecture (hereinafter ISA_{bdi}) [114]. The BDI architecture is based on Michael Bratman's theory of human practical reasoning [24]. There are other possibilities for the abstract model of the intelligent software agent, such as neural networks, but from the conceptual point of view they are not as well structured as BDI for the purpose of this thesis.

We introduce a formal Belief Description Framework (hereinafter BDF) model using an UML[3] representation. The main role of the model is to act as a bridge between the epistemological theories and an implementation. The implementation model will describe a basic architecture, which provides methods to operate on beliefs^p, justified beliefs^p, and knowledge^p. We use UML because it is widely used offering a graphical model that enables different views of a system. And it has become a de–facto standard modelling language for software engineering. UML has good extension mechanisms and semantic variation possibilities, which enable creation of profiles that can be adjusted to the purposes of various applications. A required vocabulary can be added directly into a model through the definition of classes, methods, attributes, and states.

1.1 Motivation and Problem Statement

When people share propositional information with the intention also to express the level of their confidence in information, they quite often begin their statement with phrases *I/we know that ...*, *I/we (strongly) believe that ... because ...*, or *I/we believe that ...*. And based on the used phrase a receiver establishes his/her confidence in information.[4] When today's computer systems distribute propositional information, outputs are usually only propositions expressing information without any indication of the level of confidence in information. And users tend to consider distributed information to be true (knowledge^p) because we usually tend to trust computers. But, as mentioned above, in the future we may not know (or even do not care, at all) whether the source of information is a human being or an ISA; therefore, there is a need for using same concepts in the context of information exchange regardless of the source of information. Hence, there is a requirement for ISAs to categorize the epistemic quality of information in the similar way as human beings do it. We argue that epistemology provides

[3] Unified Modeling Language

[4] Of course, there are also other factors affecting the confidence level.

proper concepts for the categorization: knowledge^p, justified belief^p, and belief^p.

In the Internet there are numerous web services, social networking services, and other information distribution services, from which users—either human beings or ISAs—can obtain information. The main trustworthiness feature of these services[5] usually is that users rely on (or do not rely on) the distributors of information meaning that the distributors are who they claim to be.[6] And the users trust on whatever basis that the distributors provide them with the correct information via dependable information distribution channels. However, several incidents have indicated that this is not a satisfactory solution [47]; for example, see the scenarios in Section 2.1.1. Retrieving information from the Internet requires an epistemically virtuous use of the Internet; however, this does not guarantee that a user will acquire justified beliefs^p or knowledge^p [68]. One of the problems is that users usually trust information distributors without any real warrants supporting trustworthiness. In order for IDS to be dependable demands additional solutions.

The epistemic quality of the piece of information, whether it is knowledge^p, justified belief^p, belief^p, or information, has or at least should have, an effect on actions taken by the users of the piece of information. Therefore, the users, especially artificial epistemic agents should have an appropriate access to the epistemic quality of the piece of information, meaning that the epistemic quality should be embedded somehow in the piece of information.

Human beings have several sources of their motivation to carry out an action, some of which are subconscious; thus, information is only one of the sources of motivation, though in some cases an important one. But in the case of ISA$_{bdi}$ information is the main source of motivation to execute an action. Therefore, the epistemic quality of information has a significant role in the motivation of ISA$_{bdi}$ to select and carry out correct actions and thus being one of the most important factors in the success of ISA$_{bdi}$'s actions.

In order to analyse and synthesize the role of the above–mentioned epistemic concepts we need to explicate several issues, such as:

1. Can ISA$_{bdi}$ have beliefs^p or are beliefs^p something only for human beings? What is the role of anthropomorphism?

2. What is the role of truth in the environment of ISA$_{bdi}$? If truth has a meaningful role, then which truth theory is the proper one?

[5]This is the case at the time of writing this thesis (25th May 2020)

[6]This is usually implemented with available certification services.

3. Can ISA_{bdi} have justified beliefs^p? If so, what justifies them? In other words, what is the most appropriate justification theory in the environment where ISA_{bdi} operates?

4. Can ISA_{bdi} have knowledge^p or is knowledge^p something only for human beings? If ISA_{bdi} can have knowledge^p, then which theory of knowledge is appropriate in the environment where ISA_{bdi} operates?[7]

5. What are the sources of knowledge^p and the sources of justification for ISA_{bdi}?

6. What is the relationship between trust and knowledge^p, justified belief^p, and belief^p in the context of ISA_{bdi}?

7. What would be the role of belief^p, justified belief^p and knowledge^p in the dependability of ISA_{bdi} and DIDS? This is one of the key questions, which needs to be answered in this thesis. The roles of knowledge^p and justified belief^p are somehow heavily intermixed with the role of trustworthiness in the services provided in the Internet. What is the relationship between them? Could knowledge^p and justified belief^p provide a better approach than today's methods to achieve trustworthiness to offer more dependable information services in the Internet?

8. An important question from the viewpoint of computer science is that can belief^p, justified belief^p and knowledge^p be modelled and implemented?

1.2 Contributions

The main and original contributions of this thesis are as follows:

1. A new, epistemological approach to the dependability of ISA_{bdi} and IDS. It is based on the epistemological theories and epistemic quality of information. This is the major contribution of this thesis.

2. Better understanding about dependability issues related to the epistemic quality of information in DIDS including ideas to design and use them. We discuss this issue in Section 2.1.1 Scenarios, in Chapters 4 Belief as Dependability Factor and 6 Belief Description Framework.

[7] As we currently have a firm confidence in ISA_{bdi} having knowledge, we need to explore how to explicate current human–related knowledge theories (e.g. reliabilism, testimony) to the environments of ISA_{bdi} (if any new explication is needed).

3. Careful analyses of epistemic value, truth, trust, and trustworthiness in the joint context of ISAs and human beings. We discuss these topics in Sections 3.2 Epistemic Value, 3.3 Truth, and 3.7 Trust.

4. Enhanced definitions of justified belief^p and knowledge^p to be adapted in the joint context of ISAs and human beings. We introduce and discuss these definitions in Sections 3.4 Belief, 3.5 Justified Belief, and 3.6 Knowledge.

5. New concepts of dependability taxonomy for intelligent distributed systems. We introduce these in Chapter 5 Enhancement to Dependability Taxonomy.

6. Belief Description Framework that introduces one proposal to model the ISA$_{bdi}$'s states of belief^p, justified belief^p, and knowledge^p including how to manage different epistemic quality of information. We introduce this in Chapter 6 Belief Description Framework.

7. A simple UML model to show implementability of Belief Description Framework. We introduce this in Appendix Belief Description Framework.

1.3 Structure of Thesis

This thesis is structured into seven topics as follows: The first chapter Introduction presents the motivation, the problem statement, and the main results. The second chapter Background and Overview provides the reader with background information and an overview of the topics, such as scenarios, dependability taxonomy, and logical issues of knowledge^p and belief^p related to ISA$_{bdi}$.

The third chapter Six Concepts examines the epistemological concepts in the context of ISA$_{bdi}$ and introduces an approach to the definitions of truth, belief^p, justified belief^p, and knowledge^p. It also discusses trust and trustworthiness to explicate them in the context of ISA$_{bdi}$.

The fourth chapter Beliefs as Dependability Factors introduces beliefs^p, justification and justified beliefs^p, and knowledge^p as dependability factors. It also discusses some major problems with implementing belief^p–related dependability.

The fifth chapter Enhancements to Dependability Taxonomy introduces required enhancements of the dependability taxonomy.

The sixth chapter Belief Description Framework presents the model of a framework to represent, manage, and distribute knowledge^p, justified belief^p, and belief^p.

The seventh chapter Summary presents the summary of the results of this thesis.

Chapter 2

Background and Overview

2.1 Introduction to Dependability Issues

In this section we introduce scenarios which are used to illustrate our motivations and problems related to the epistemic quality of information. We also use the scenarios to evaluate of our solutions. The scenarios deal with issues such as untrue tweet, the correctness of diagnoses, and the dependability of a traffic information service. We also present the part of Jean–Claude Laprie's et. al. dependability theory that is relevant to this thesis.

2.1.1 Scenarios

In this section we introduce and discuss three illustrative scenarios that will attract attention to the importance of belief^p, justified belief^p, and knowledge^p in the context of dependable IDS. The first scenario discusses knowledge^p, justified belief^p, and belief^p and their significance in a social media. The second scenario presents an emergency medical case, where belief^p, justified belief^p, and knowledge^p play significant roles in the proper treatment of a patient. The third scenario examines a traffic information service which illustrates some of the implementation issues of our Belief Description Framework.

When discussing these scenarios we assume proper justification and knowledge theories to be forms of reliabilism[1], and testimony[2] to be as

[1] Alvin I. Goldman: *"If S's belief in p at t results from a reliable cognitive process, and there is no reliable or conditionally reliable process available to S, which had it been used by S in addition to the process actually used, would have resulted in S's not believing p at t, then S's belief in p at t is justified."*

[2] Jennifer Lackey: *"For every speaker S and hearer H, H comes to know that p via S's*

a transfer method of justified belief^p and knowledge^p. In Chapter 3 we motivate this assumption.

Faked White House Bomb Tweet Causes Stock Market Panic

On the 23rd of April, 2013, at 13:07, the following tweet was delivered from the Associated Press [36]: "Breaking: Two Explosions in the White House and Barack Obama Injured."

The stock market reacted immediately. The Dow Jones fell by in a matter of seconds about 140 points, which is more than a full per cent of its value. When it had become clear that the tweet was not true, the Dow Jones regained almost everything it had lost within 10 minutes of the untrue tweet.

It turned out that the Twitter[3] account of the Associate Press had been cracked.

Reports suggest more than 20 billion dollars worth of equity positions changed hands on the New York Stock Exchange during the brief trading hiccup.

Thus, some traders made big profits, and some traders made significant losses within those 10 minutes. Therefore, we are entitled to raise several questions: Why did this happen? Why did traders on Wall Street not collide with social media, when a false tweet from a trusted source was distributed? Why did traders rely on this piece of information? Is Twitter trustworthy? Is The Associated Press[4] trustworthy? Do not traders care? Can this kind of incident be avoided in the future by having more trustworthy social media services? In this thesis we address some of these questions.

This scenario points out an attitude of trusting without any specific formal warrant for information on a well–known information distributor to distribute only news that are true.[5] In the future there will be more and more automatically generated—written by AI–based applications—news,[6] and therefore, this kind of attitude is no longer acceptable. There will be

statement that p only if (i) S's statement that p is appropriately connected with the fact that p; (ii) H has no defeaters indicating the contrary."

[3] www.twitter.com

[4] https://www.ap.org

[5] The Associated Press is one of the oldest news agents and considered to be trustworthy.

[6] An interview with Professor Kristian Hammond by Steven Levy in Wired Magazine; see url www.wired.com/2012/04/can-an-algorithm-write-a-better-news-story-than-a-human-reporter/all/

a requirement to provide some kind of a warrant of the epistemic quality associated with news.

We can consider the tweet *"Breaking: Two Explosions in the White House and Barack Obama Injured."* to be a combination of three propositions expressed in "tweet language"[7]. The propositions are as follows:

1^{st} proposition: *Breaking*: This is a breaking news item.

2^{nd} proposition: *Two Explosions in the White House*: There have been two explosions in the White House.

3^{rd} proposition: *Barack Obama Injured*: President Barack Obama is injured.

The logical expression of the tweet is the following one: *"this is a breaking news item* and *there have been two explosion in the White House* and *President Barack Obama is injured"*.

None of the beliefs^p (propositional attitudes) based on these propositions is the result of reliable^p cognitive processes. In this case there are two main cognitive processes involved in the belief^p–forming. The first one is the cracker's process of the proposition creation. Our intuition claims that the process that deliberately results in lies is not reliable^p (*reliabilism*). The second one is the belief^p–forming process of the receiver. Even though the process itself could be reliable^p, the receiver of the tweet cannot come to know the beliefs^p as they are not appropriately connected with the facts (*testimony*). Therefore, we can claim that there is no justification^p for the beliefs^p and the beliefs^p are not knowledge.

When we evaluate the beliefs^p from the receiver's subjective viewpoint, the outcome seems to be different. The receiver considers that both his belief^p–forming process and the process, which the Associated Press uses to publish tweets, are reliable^p enough. The Associated Press mostly produces reliable^p news, and it cannot be cracked. And for the first ten minutes after the tweet there is no reliable^p or conditionally reliable^p process available to the receiver that would result in the receiver not to believe the propositions. Therefore, the receiver's beliefs^p for the first ten minutes are justified (*reliabilism*). But his/her beliefs^p are not knowledge^p, as the beliefs^p are not true; though the receiver is not aware of it.

There is an obvious demand for some kind of a certification service that classifies news to different categories according to their epistemic quality (trustworthiness); for example, to be either information, belief^p, justified

[7]We see the tweet language to be a kind of short–hand expressions, which is due to the limitation of the Twitter service.

belief^p, or knowledge^p. And the category should be embedded with news in order to allow a receiver to evaluate the usefulness of news.

Traders on Wall Street seem to be too tightly intertwined with Twitter, an information service, which acts as both a gossip distribution media and news service, where individuals, professional journalists, and publishers send out breaking news. Traders seem to rely on information, which is not certified as either knowledge^p or justified belief^p. The warrant service would provide traders with a better possibility to evaluate news, and therefore to achieve overall better results.

A Tourist having an Accident in a Foreign Country

The following scenario[8] illustrates the progress of diagnoses from belief^p to knowledge^p. The entity, which we are discussing in this scenario, is the diagnoses of trauma, which are expressed as propositions *"The correct diagnosis is ..."*. These change along getting more reliable^p information.

Phase 1 — An accident and the first diagnosis

Mr. Matti Meikäläinen, a 30–year–old action actor, spent his vacation in a small town in Thailand. Having spent five relaxing days in the town he decided to see also the surrounding countryside. He rented a motor scooter and started to drive towards a nearby small fishing village. Unfortunately, Matti was not used to the left–hand traffic, and therefore, he had a traffic accident in a road crossing. Matti's right ankle was stuck under the motor scooter when it fell over. It seemed to cause a low–energy trauma, as luckily Matti was driving slowly. Matti drove slowly back to the hotel and went directly to visit a nearby nurse, whom a person at the hotel's reception desk recommended. The nurse checked Matti's ankle and said that it was not serious, just some muscles were sprained. However, next morning Matti's ankle was really painful and swollen, and he could not step on his foot. Matti was transferred to the district hospital. While waiting for X–ray image to be taken, Matti activated his Travellers' Health and Insurance Service (hereinafter THIS) application, which was installed on his smart phone, and he typed in, following the instructions given by THIS, all the details of his accident. THIS application contacted Matti's travel insurance company and sent all the details to the company. An on–duty physician at the district hospital analysed the X–ray image and came to the conclusion that there is a lateral malleolus fracture in Matti's ankle. The

[8] We developed this hypothetical but quite possible scenario together with Doctor Ari Kinnunen, who was the co–founder and the medical director of EMA (Emergency Medical Assistant) Group.

physician said to Matti: ***"The correct diagnosis is lateral malleolus fracture"****. Based on the diagnosis an orthopaedic cast was laid to stabilize the ankle.*

Phase 2 — Further examinations and the second diagnosis

Matti's travel insurance company granted EMA (Emergency Medical Assistant) in Finland to carry out the care monitoring and other required actions to ensure the best possible medical care for Matti. The Travellers' Medical Assistance (hereinafter TMA) application of EMA retrieves Matti's relevant medical history from the National Health Archive of Finland (www.kanta.fi) in order to verify that Matti does not have any such illnesses that must be taken into account in Matti's treatment. There were no such illnesses. The proper functioning of the ankle is essential in Matti's profession; therefore, a physician at EMA requested via TMA for a copy of the X–ray image from the hospital. TMA received the copy and carried out a first–level analysis, the result of which indicated that the X–ray image was low quality, one axis image. TMA displayed the X–ray image and the data of its quality to the physician on duty, who realized that the X–ray image may not have revealed all the possible fractures because of its low quality. She requested via TMA that Matti must be transferred to a hospital having facilities for a higher quality X–ray imaging. TMA organized the transfer together with local people in Thailand. Matti was transferred to a private hospital in Bangkok, where other—this time a high quality, multi–axes—X–ray images were taken. TMA retrieved the new X–ray images from the Bangkok hospital, and sent them to a consulting Finnish radiology center specialized in detecting even minor fractures, which are usually difficult to observe in X–ray images. The center uses a new computer–aided diagnosis (hereinafter CADx) system to interpret X–ray images automatically. The CADx system found out from the X–ray images that the previous diagnose was not correct, but there was a bimalleolus fracture, which could cause a permanent ankle disability without proper operation. The CADx system stated: ***"The correct diagnosis is bimalleolus fracture"****, and the reliability of the diagnoses is based on the high quality image scanning, the reliability of which is 0.999. The CADx system sent the interpretation to the TMA application of EMA, which informed the physician on duty of the diagnosis. TMA also informed Matti via his THIS application about the new diagnosis. The wrong diagnosis could have ended Matti's career as an action actor.*

Phase 3 — Operation and the third, final diagnosis

The physician at EMA decided to transfer Matti back to Finland in or-

der for the ankle to be operated and for the proper post–operative care. A nurse was sent to Bangkok to escort Matti back to Finland because Matti had a high risk of deep venous thromboses, the prevention of which required low molecular heparine medication. The nurse escorted Matti to Helsinki University Hospital, where Matti's ankle was operated. The operation revealed that there in fact were trimalleolus fractures, and screws and a plate were required to be placed to support the normal alignment. The orthopaedist verified that ***"The correct diagnosis is trimalleolus fracture"****. The operation and proper post–operative care shortened Matti's recovery significantly and prevented the permanent disability of Matti's ankle.*

This scenario indicates the importance of comprehending the differences between belief^p, justified belief^p, and knowledge^p to the success of the medical care. If the medical care would have been carried out only on the basis of the belief^p without proper justification (required level of reliability^p), it could have resulted in permanent disability and an unnecessary spending of health care cost. We analyse this scenario in more detail in Sections 2.2.2 and 3.2.

Traffic Information Service

The following scenario of traffic information service[9] (hereinafter TIS) is used to demonstrate issues in the defining of the required reliability^p of belief^p, justified belief^p, and knowledge^p. It is also used to demonstrate the scheme of possible worlds using an example of the ontology of TIS.[10]

The scenario is as follows: *In the environment of Road 101 there is a traffic information service that informs the drivers of approaching vehicles about the driving conditions on Road 101. There are three declarations of the driving conditions: (1) When the road might be slippery a notice is displayed. (2) When there are clear indications of the road being slippery a warning is displayed. (3) When it is certain that the road is dangerously slippery an alert is displayed. TIS is provided in co–operation by several ISA_{bdi}s and human beings. The role of ISA_{bdi}–A is to announce traffic notices, warnings, or alerts both to human drivers and autonomous vehicles driven by ISA_{bdi}s, when vehicles are approaching Road 101, and the belief^p of ISA_{bdi}–A* ***"Road 101 is slippery."*** *fulfils specified epistemic requirements.* TIS is illustrated in Figures 2.1 and 2.2.

Let us have an example of the processes of TIS (Figure 2.3). We assume that ISA_{bdi}–A perceives from a source X the proposition *"Road 101*

[9] This is purely a hypothetical example in order to clarify our thinking about the roles and sources of information in DIDS.

[10] See Appendix *Discussions on Evaluating Epistemic Quality of Beliefs.*

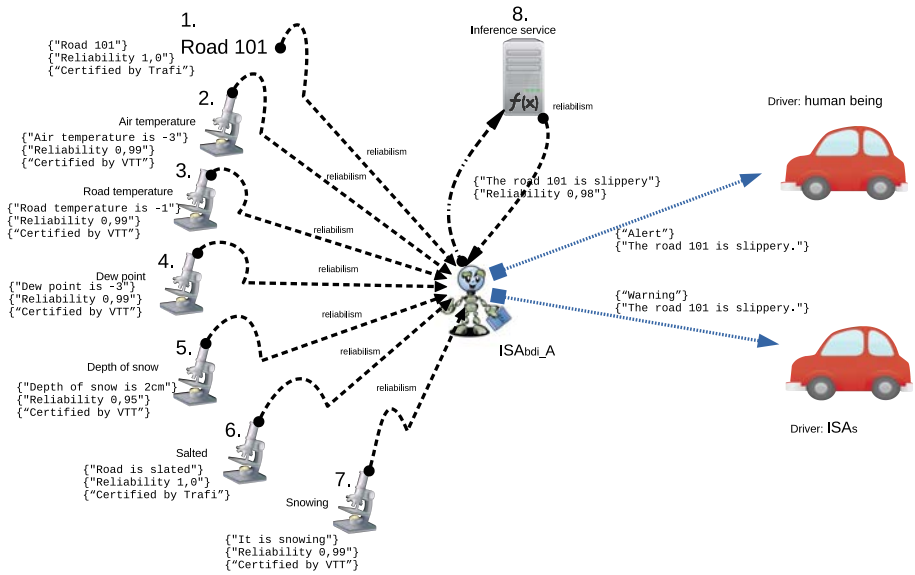

Figure 2.1: A scenario of traffic information service.

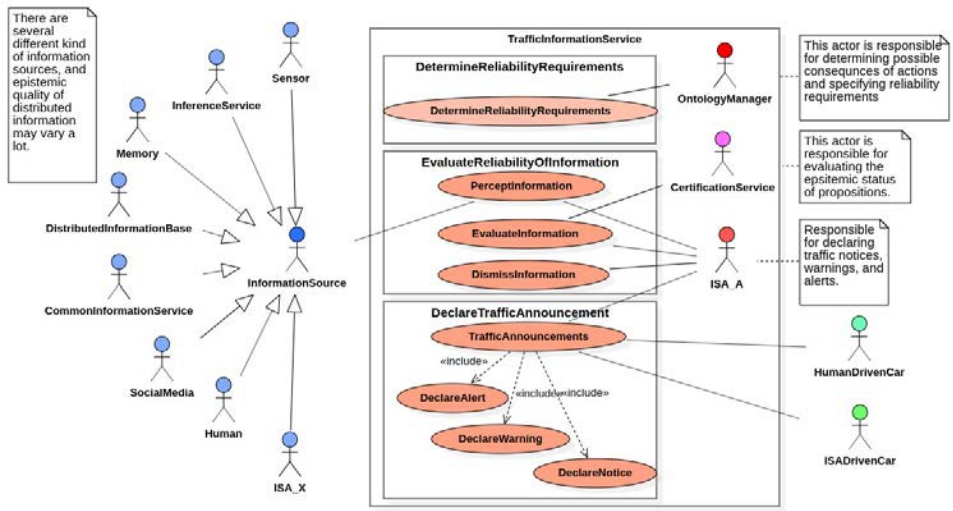

Figure 2.2: UML use case of traffic information service.

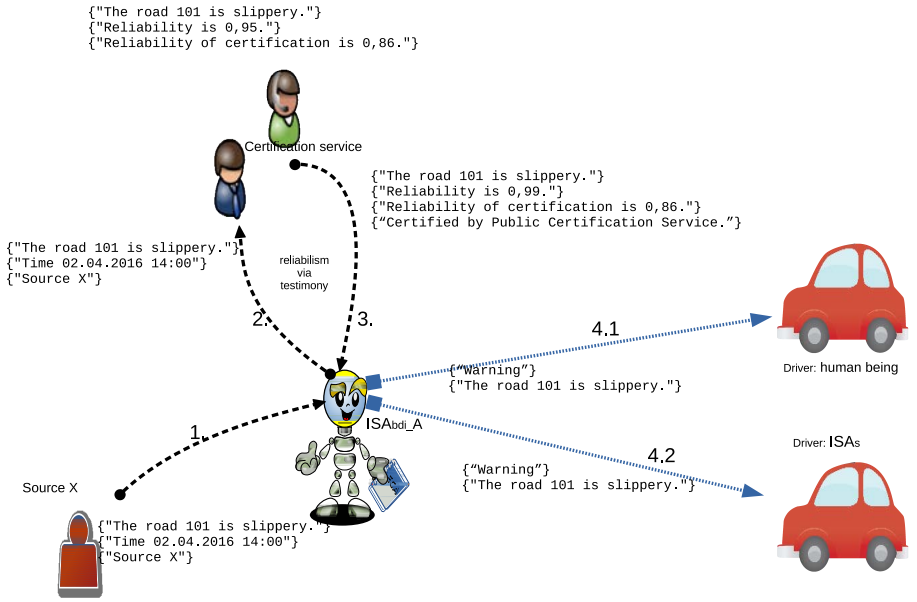

Figure 2.3: An example of TIS utilizing a certification service.

is slippery." including the metadata *"Time 02.04.2016 14:00" "Source X"*. Because there is no reliability^p data of the creation process of the proposition available, ISA$_{bdi}$–A sends the proposition to a certification service in order to get the certificate of the epistemic quality of information. Let us further assume that after the evaluation ISA$_{bdi}$–A perceives from the certification service: *"Road 101 is slippery." "Reliability^p is 0.95." "Reliability^p of certification is 0.86." Certified by Public Certification Service.* The third one expresses the reliability^p of the certificate creation process of the certification service. Based on this ISA$_{bdi}$–A forms the belief^p *"Road 101 is slippery."* with the associated metadata. There are two separate factors to be taken into account when inferring whether or not to announce a traffic notice, warning, or alert. In this case the reliability^p does not fulfil the requirement for the belief^p to be knowledge^p, as the reliability^p of the certification process is not high enough. But it is high enough for the belief^p *"Road 101 is slippery."* to be justified belief^p. Therefore, ISA$_{bdi}$–A declares the traffic warning both to ADC and to HDC.

The scenario of TIS is used and further discussed in more detail in Section 4.2.2 and in Chapter 6.

2.1.2 Dependability Theory

We commonly characterize computing systems with the following properties: functionality, usability, performance, dependability, adaptability, manageability, and cost. Since the first generation of digital computers the dependability[11] of computer systems has been an important topic of computer science. Early computers were built using unreliable^c components, therefore, research on dependability started with developing practical techniques to improve their reliability^c. As an example we can mention the redundancy theories that C.E. Shannon, J. von Neumann, and E. F. Moore developed [97]. In the decades of 1980 and 1990 Jean–Claude Laprie et. al. developed a consistent set of the concepts and terminology of dependability and published them in the book *Dependability: Basic Concepts and Terminology* [82].

We argue that the latest developments in the domains of AI, ISAs and autonomous robots change the scene in such a way that the dependability concepts and terminology need to be enhanced to take into account the effects of learning, autonomous operation, and varying epistemic quality of information. For example, the current dependability taxonomy does not properly address environments, where ISA$_{bdi}$—or a robot—operates with uncertain information (not knowledge^p) or learns by the trial–and–error method. We address these problems below and in Chapter 5 (Enhancement to Dependability Taxonomy).

Basic Concepts and Taxonomy

We can look at dependability from two different viewpoints: we emphasise either qualitative factors or quantitative factors. We can consider the dependability of a system to be either the ability to deliver service that can justifiably be trusted or the ability to avoid service failures that are more frequent and more severe than is acceptable to the users [11]. The former viewpoint begs the question of what does *"justifiably be trusted"* actually mean. We will discuss justification in Section 3.5 and trust in Section 3.7 from the philosophical viewpoint. The latter one is more straightforward from the viewpoint of computer science because the concept *"more frequent and more severe than is acceptable to the users"* is easier to actualise, for example, by measurements in usability tests or system acceptance tests [12]. There is a causal relationship between these two definitions: we commonly obtain justification for trust when there are less service failures and service failures are less severe than we are willing to accept.

[11] Mostly called reliability^c at that time.

There are other definitions of dependability—usually established for special application domains—such as follows: *"The collective term used to describe the availability performance and its influencing factors: reliability performance, maintainability performance, and maintenance support performance"* [102] and *"The extent to which the system can be relied upon to perform exclusively and correctly the system task(s) under defined operational and environmental conditions over a defined period of time, or at a given instant of time"* [73].

The dependence of an entity A on another entity B represents the extent to which A's dependability is affected by that of B. Trust is accepted dependence. The relation *depend upon* is defined as follows: A depends upon B if the correctness of B's service delivery is necessary for the correctness of A's service delivery. Accepted dependence is about the judgement that this level of dependence is acceptable.

The basic concepts of the dependability taxonomy comprise the following terms [11]:

1. A *system* is an entity that interacts with other entities, i.e., other systems, which form the environment of the given system.

2. A *system boundary* is the frontier between the system and its environment.

3. The *function* of a system is what the system is intended (described by functional specifications) to do.

4. The *functional specification* of a system describes what the system is intended to do in terms of functionality and performance.

5. The *behaviour* of a system is what the system does to implement its function. The behaviour is described by a sequence of states of the system.

6. The *total state* of a system comprises the following states: computation, stored information, interconnection, and physical condition.

7. The *structure* of a system enables the system to generate its behaviour.

8. The *service* of a system is the behaviour of the system as it is perceived by its users.

9. A system delivers *correct service* when the service fulfils the system function.

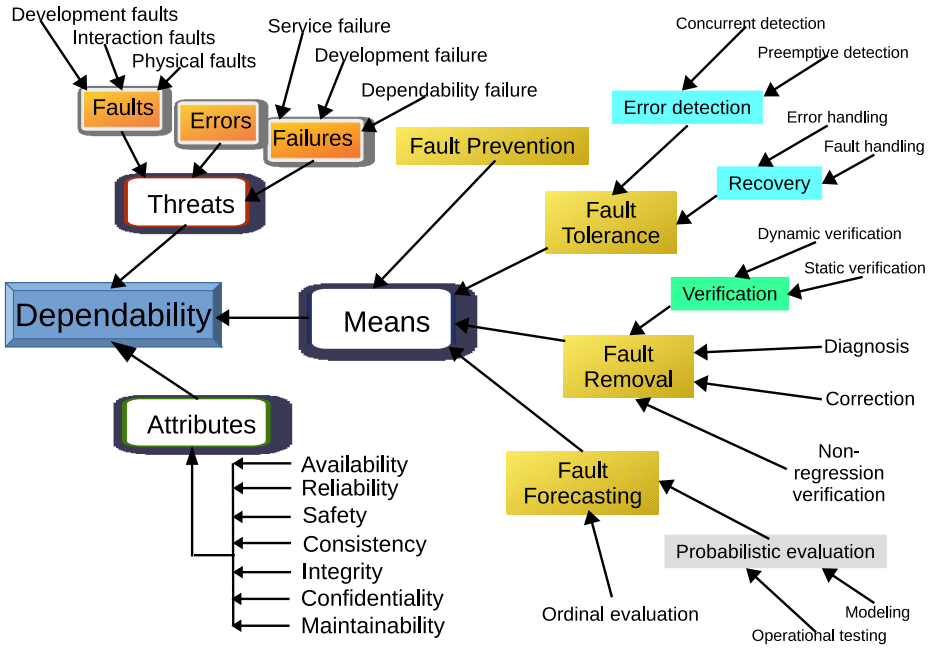

Figure 2.4: Dependability taxonomy.

10. A *service failure* is an event that takes place when the delivered service deviates from the correct service.

11. A *service outage* is the period of the delivery of an incorrect service. *Service failure modes* are ranked based on *failure severities*.

12. A *degraded mode* of system exists, when the system is capable to offer only a subset of the needed services.

13. The *external state* of system is the part of the total state of the system that is perceivable at the service interface.

14. The *internal state* of system is the part of the total state of the system that is not perceivable at the service interface.

Jean–Claude Laprie et.al. model dependability as illustrated in Figure 2.4 [11, 12, 83]. The dependability taxonomy comprises three sets of factors that are attributes, impairments, and means. The attributes are the following ones:

1. Availability is the readiness for usage.

2. Reliability[c] is the continuity of service.
3. Maintainability is the ability to undergo repairs and evolution.
4. Confidentiality is the non–occurrence of unauthorized disclosure of information.
5. Integrity is the non–occurrence of improper alterations of information.
6. Consistency is the logical coherence of data or the logical coherence of co–operating processes.
7. Safety[c] is the non–occurrence of catastrophic consequences on the environment.

There exist also secondary attributes such as the following ones:

1. Accountability: availability and integrity of the identity of the person that performed an operation.
2. Authenticity: integrity of the content and origin of a message, possibly of some other information, such as the time of emission.
3. Nonrepudiability: availability and integrity of the identity of the sender of a message.

The impairments are as follows:

1. Faults are the causes of errors.
2. Errors are the deviations from the correct service states.
3. Failures mean that at least one (or more) external state of the system deviates from the correct service state.

The development of a dependable computing system requires a combined set of methods and techniques:

1. Fault prevention: means to prevent fault occurrence or introduction.
2. Fault tolerance: means to ensure that a service fulfils the function of the system in the presence of faults.
3. Fault removal: means to reduce the presence of faults.
4. Fault forecasting: means to estimate the present number, the future incidence, and the consequences of faults.

The core features of Laprie's dependability model are based on the assumption that dependability is a technical attribute and the dependable features are within the computing systems themselves. The model has the following assumptions as its guidelines [32]:

- Errors arise inevitably from faults.
- A system is constructed so that an error could be detected by an external observer.
- Users are able to recognize the occurrences of system failures.

We claim that the above assumptions will not hold in the future. This taxonomy of system dependability needs to be enhanced in order to be applicable in the environment of future dependable intelligent distributed computing systems based on AI, ISA, and robots. The role of computing systems in the society is rapidly changing towards autonomous agents, which are operating increasingly in a social environment of uncertain information. Therefore, the importance of recognizing whether information is belief^p, justified belief^p or knowledge^p and acting based on the epistemic quality of information increases in the determination of the dependability of ISA and IDS. There are also other domains, such as Advanced Persistent Threats [29] and dependability of cyber–physical systems [124], which have also addresses the need for enhancements to the dependability taxonomy.

2.2 Intelligent Distributed Systems

In this section we discuss some features of intelligent distributed systems. We define a system to be an intelligent distributed system as follows:

Definition. An intelligent distributed system is a collection of independent agents that appears to its users as a single coherent system, where an independent agent can be either an intelligent software agent, a robot, a process running in a computer, or a human being, and some of the independent agents are software–based entities, of which some are implemented utilizing artificial intelligence.

An example of an intelligent distributed system is illustrated in Figure 2.5, where a single coherent system providing a service to users is built up by several independent agents, such as an inference system, a distributed information base, intelligent software agents, social media, a professional

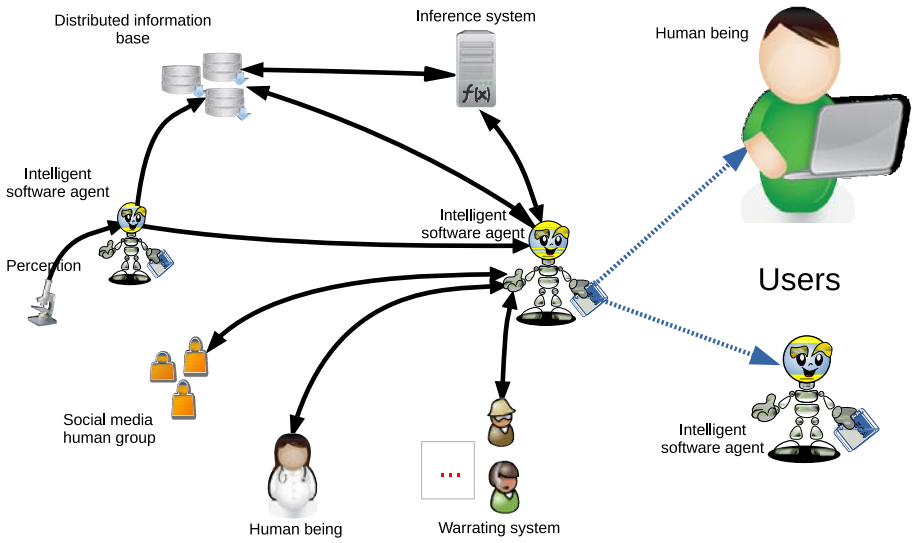

Figure 2.5: An example of an intelligent distributed system.

human being, and an information certification service. A user can be either a human being or an intelligent software agent acting as an *epistemic agent*. We define the epistemic agent as follows:

Definition. An epistemic agent is an entity (either a human being or an intelligent software agent) that has an important effect on a situation and perceives, holds, processes, and distributes semantical information.

At first we discuss briefly the main features of AI that are relevant to this thesis, such as GOFAI (Good Old Fashioned AI), connectionist models (a.k.a. neural networks and deep learning), ISA_{bdi}, and representations of semantic information.[12] Then we proceed to discuss the role of knowledge^p, justified belief^p, and belief^p in DIDS. Finally, we discuss logical issues related to belief^p, justified belief^p, and knowledge^p.

[12]The actual topic of this thesis is not AI, but features that are required in AI–based solutions

2.2.1 Artificial Intelligence

In this section we discuss the following areas of AI: GOFAI, connectionist models, intelligent software agents, and representations of semantic information.

AI is an approach consisting of many disciplines to understand, model, and implement intelligence and cognitive processes. Tools such as mathematics, logic, computation, and mechanics are used to realize AI. Philosophy has had a significant role in AI because the concept of truth has been important in both AI research and epistemology; foundational questions of AI are philosophical in nature; and philosophical concepts, such as knowledge, information representation, and action need to be understood properly in AI in order to model and implement them. On the other hand, AI raises new questions in metaphysics, ethics, and epistemology, such as how intelligent behaviour ought to be explained or how to understand human intelligence.

AI comprises several themes such as smart software versus cognitive modelling, symbolic AI versus connectionism (a.k.a. neural networks or deep learning), reasoning versus perception, reasoning versus knowledge, to present or not to present, and narrow AI versus human–level intelligence [42, 92]. In this thesis we work on a cognitive modelling to establish a model for ISA to have information, belief^p, justified^p, and knowledge^p. We concentrate on symbolic AI because it better provides an environment where information can be classified based on the epistemic quality, and propositions are presented symbolically by nature. In the case of reasoning versus perception our approach is more close to perception than reasoning. And also in the case of reasoning versus knowledge^p we concentrate on knowledge^p because in the real world systems with a significant amount of information we must know and model the epistemic quality of information. In the case of to present or not to present we argue that a system shall model its world, at least to the amount, where possible consequences of an action can be evaluated to a required dependability. We do not have any strong opinion about narrow AI (weak AI) versus human–level intelligence (strong AI) despite the fact that we argue that ISA is capable to have belief^p, justified belief^p, and knowledge^p.

GOFAI

GOFAI is a label that denotes classical, symbolic AI [15]. The basic idea of GOFAI is to operate on programmed instructions and formal symbolic representations. GOFAI symbols and programs composed of them are re-

garded as purely formal—having semantics—structures, and GOFAI computation involves the construction and transformation of symbolic data structures [15]. For example, a proposition and a program evaluating the epistemic quality of propositional information can be implemented using GOFAI. Most of the intelligence in GOFAI lies in the choices of actions and heuristics specified by the programmer. GOFAI programs often simulate the conscious deliberation of high–level human thought because in GOFAI propositions are presented with specific semantic content. One strength of GOFAI is the ability of representing propositional contents. We claim that the epistemic quality of input information plays an important role, when a GOFAI program in a running state decides that a particular action is needed to achieve its goal.

The frame problem plays a role in GOFAI. There are two aspects: first, knowing which factors in a situation would be changed by an action and which would not, and second—the more important issue in this thesis—reasoning with incomplete information due to the variety of the epistemic quality of information in the real world and the vagueness of ordinary language concepts. The frame problem is probably insoluble for the general case, but it has been and will be solved for specific purposes in many different environments [15].

The key strengths of GOFAI are modelling multi–level hierarchy, sequential order, and inferential relations between specific propositional contents [15].

Connectionism

Connectionism is the current, dominant domain of AI. Connectionism is the way of capturing and understanding the mechanisms and processes of cognition through building models using networks of simple, neuron–like processing elements, each of which perform simple numerical computation. The main idea of connectionism is that cognition is a result of an interaction of a large number of simple processing units (i.e., the large number of connected neurons in the brain). A representation is a pattern of activation over a set of processing units in a model. Processing is carried out through the propagation of activations among the processing units and via the interconnections among them. The propagation of activations is mediated by the numerical connection weights between pairs of processing units. Learning takes place through the change of the connection weights. According to connectionism cognition should be approached more in terms of mechanisms of constraint satisfaction, pattern recognition, and weight adaptation, rather than explicit symbol manipulation [144].

Connectionist representations can be categorized into two main categories: localist representations and distributed representations. In localist representations each node represents a single concept, and in distributed representations each concept is represented by an activation pattern over a set of nodes [144]. Memory is often a constructive process involving the interactions of simple processing units. Because the representation of information is not in a linguistic form, connectionism causes problems in the evaluation and representation of the epistemic quality of information. In localist representations the problems are not so severe as in distributed representations, because each unit is interpretable and has a clear conceptual meaning. Each unit also captures the property of explicit information; thus, information being better accessible and more manipulable [144].

Connectionist models face difficulties when higher–level cognition, such as reasoning and problem solving, is required as well as in the case of the binding problem (the combination of multiple arbitrary in processing and representation) [144].

Hybrid Approach

Because of the problems of pure symbolic and pure connectionist models, hybrid models have been proposed to resolve the problems [144]. Symbolic models work better in the domains of search and knowledge representation. Search comprises domains such as a systematic exploration of a space of problem states and a means of conceptualizing and conducting problem solving. Knowledge representations comprise domains such as logic–based representations, structured representations (i.e. semantic networks), and production rules. Connectionist models work better in the domains of implicit information, learning, parallelism, and reasoning by constraint satisfaction and pattern recognition. Hybrid models tend to combine the best features of both approaches. The result would be more expressive, more powerful, often more efficient, and more useful in both cognitive modelling and practical applications, as cognitive processes are not homogeneous. Cognitive processes consist of a wide variety of information representations and processes that play different roles and serve different purposes.

An architecture incorporating both symbolicist and connectionist models can be implemented computationally by a combination of a symbolic system (explicit information) and a connectionist system (implicit information). We can classify at a high level hybrid models as follows:

1. Closed, meaning that a system comprises explicit information or information can be inferred from explicit information within an accept-

able time–frame. Closed models with formal symbolic representations are suitable to be implemented using symbolic models.

2. Semi–open, meaning that a system comprises, in addition to explicit information, information that cannot be inferred within the required time–frame. Semi–open models are suitable to be implemented using hybrid models.

3. Open, meaning that a system comprises mainly implicit information. Open models are suitable to be implemented using connectionist models.

We address in this thesis closed and semi–open models.

Hybrid models can lead to complicated architectures and systems because they may comprise a variety of different types of processes and representations, and multiple heterogeneous mechanisms interacting in a complex way. Therefore, the following issues are raised [144]: First, how to decide which representation (symbolic, localist, or distributed) is most proper for each part of the system. Second, how do learning and symbolic representation interact? Third, how do we structure different parts of a hybrid system to achieve optimal results? Fourth, how can complex symbolic structures (rules, frames, and semantic networks) be learned?

We are of the opinion that hybrid models, where the epistemic quality of information is managed by the symbolic models, are the most proper ones for many contexts and environments of ISA and DIDS.

Intelligent Software Agents

The development of software technology and AI, in particular, in the last two decades has established a new foundation for software–based systems. These systems, which are generally called agent–based systems, quite often act independently on behalf of human being, and they demonstrate more and more human–like behaviour. In the domain of IDS artificial epistemic agents can be implemented using a paradigm called intelligent software agent technology. There is no unambiguous definition of the term intelligent software agent. However, a descriptive one is as follows: *An intelligent software agent is a computational entity that can be viewed as perceiving and acting upon its environment and that is autonomous in that its behaviour at least partially depends on its own experience* [162]. There are different types of ISAs, such as collaborative agents, personal agents, information agents, and various combinations of these agents. A number of various

studies of ISAs have resulted in different kinds of models, such as reactive, goal–directed, and deliberative models.

There has been several approaches to capture the idea of ISA. The approach of *knowledge and action* (Robert C. Moore) concentrated on the question of what an agent needs to know in order to be capable of performing an action [166]. The approach of *intention* (Philip R. Cohen and Hector J. Levesque) studied the concept of *intending to act* that defines conditions for an agent to perform an action. This approach used two basic attitudes that were belief^c and goal [166]. This introduced belief^c as one of the basic notions of the intelligent software agent theory. In related work Anand Rao and Michael Georgeff developed a model based on beliefs^c, desires^c, and intentions^c that resulted in a *belief–desire–intention* (hereinafter BDI) architecture for the internal structure of ISA [26, 113, 114]. BDI is based on the ideas of Michael Bratman's philosophical theory of practical reasoning [24]. Beliefs^c represent characteristics of an environment, which ISA$_{bdi}$ perceives whenever needed. Desires^c represent goals to be achieved as well as properties associated with goals. Intentions^c represent selected actions to achieve a desired goal.

The BDI model is the most suitable one for the objectives of this thesis. First, it is based on the idea of modelling the activity (practical reasoning) of human being. Second, the concept of belief^c is adequately similar to the one of epistemology, so that it can be used to discuss and define belief^p, justified belief^p, and knowledge^p for the joint environment of human beings and ISAs. Third, beliefs^c and desires^c can be represented as states comprising propositions. And finally, BDI is the widely accepted model. Therefore, in this thesis we utilize ISA$_{bdi}$ as a high level, abstract model of artificial epistemic agents of IDS that processes, evaluates, and manages information and its epistemic quality.

Typically, ISA$_{bdi}$ has a representation of the state of the world and a representation of the desired state of the world. As beliefs^c contain information (propositions) that ISA$_{bdi}$ has about its surrounding world, it is the entity that we discuss in this section. ISA$_{bdi}$'s sources of beliefs^c (knowledge^p, justified beliefs^p and belief^p) can be, for example, perception, introspection, memory, reason, and testimony. ISA$_{bdi}$'s perceptual capabilities can be implemented with various kinds of input devices, such as video/infrared camera and radar (sight), microphone (hearing), pressure sensor (touch), air flow sensor (smelling) (in general, different kinds of environmental sensors), keyboard, touch screen, etc. Introspection can be thought to be ISA$_{bdi}$'s capacity to inspect its internal state: which beliefs^c are stored in ISA$_{bdi}$'s memory, what is the amount of beliefs^c, what is the

maximum capacity to hold beliefs^c, what is the status of beliefs^c, what is the status of inferring processes, etc. ISA$_{bdi}$'s memory comprises semantic, structured data—actual data and associated semantic metadata[13]—that can be stored, for example, in a main memory, in an external disk, or in a cloud. Reasoning establishes—in addition to perceiving—the important source of beliefs^c. It is also most studied subject in the domain of ISA$_{bdi}$; but nevertheless, it is still a very difficult issue. Actual knowledge does not seem to obey any logic [34]. And, in addition, it is not yet possible to build an ISA$_{bdi}$ that possesses the reasoning power described by normal modal systems [154]. However, various modal epistemic logic has been proposed, for example, one by Ho Ngoc Duc [34] and another one by Michael Wooldridge [165]. Testimony is considered to be the source of beliefs^c when ISA$_{bdi}$ acquires propositions from other ISA$_{bdi}$s or human beings.

As already discussed above, ISA$_{bdi}$ exhibits human–like behaviour, but what kind? There is no unambiguous definition about what kinds of behavioural properties ISA$_{bdi}$ should have. ISA$_{bdi}$ usually exhibits the following prominent properties:

1. Autonomy: ISA$_{bdi}$ acts independently on behalf of its master without human intervention.

2. Proactive: ISA$_{bdi}$ is capable to create or control a situation by causing something to happen rather than responding to it after it has happened.

3. Goal–oriented: ISA$_{bdi}$ has its own desires which it tries to achieve.

4. Collaborative: ISA$_{bdi}$ is capable of working jointly with other ISAs and/or human beings on activities.

5. Communicative: ISA$_{bdi}$ is capable of exchanging beliefs^c (knowledge^p, justified beliefs^p, and beliefs^p) with other ISA$_{bdi}$s and/or human beings.

Figure 2.6 illustrates one possible BDI architecture of ISA$_{bdi}$. In this BDI architecture most interesting components are perception as the interface to the external world, beliefs^c (world model, mental model, and social model) as propositions about the external world, and situations (routine emergency, local planning, and co–operating) as propositions about the internal state. The belief generation process is the activity that would deal with the epistemic evaluations of beliefs^c.

[13] This data represents propositions about the external world.

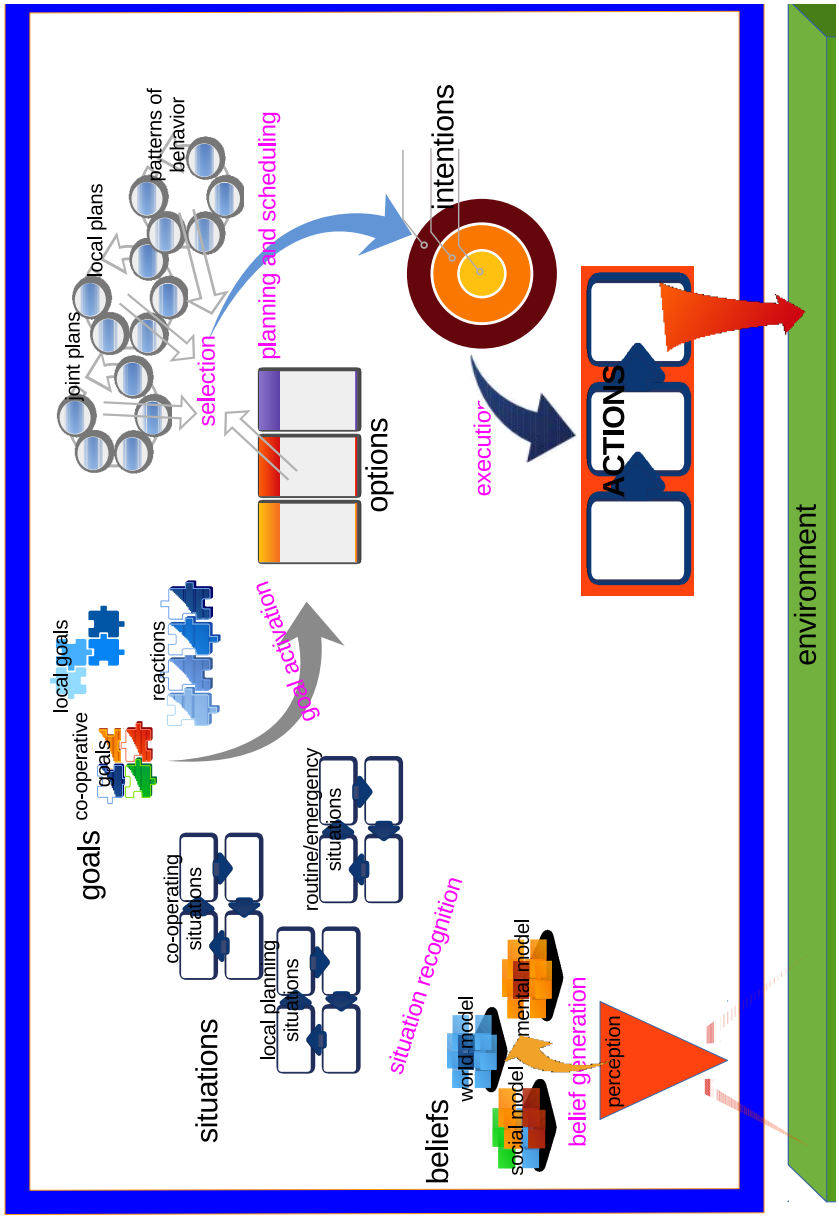

Figure 2.6: BDI architecture.

An overview of the BDI–based reasoning process is as follows:

1st Perceive the environment to generate beliefs^{c}.

2nd Choose a desire^{c} to achieve, then select a plan to reach that desire^{c}.

3rd Decide based on the plan the next action to perform.

4th Perform the selected action.

5th Every now and then check if the plan is still valid.

ISA$_{bdi}$ is in general continuously interacting with its environment. The environment is not only a source of problems for ISA$_{bdi}$ to solve, but rather a co–operation component with which ISA$_{bdi}$ is involved. From the viewpoint of justification and knowledge^{p}, perception and belief^{c} generation are the key entities in the BDI architecture.

Representation of Semantic Information

The essence of the issue of representation is that representations are critical for the process of deciding what action to take, and not so much for the process of executing the action [42]. In the context of ISA$_{bdi}$ there are two representation lines, which are logic–based and probability–based [4]. In this thesis we discuss logic–basic representations as they are more suitable to represent propositions; despite of the fact that some information types, such as spatial, temporal, and uncertain information are difficult to represent using a sentential language. There are two domains: first, representing the world (ontology^{c}) where information is used, and second, representing information (propositions about the world). The representation of semantical information affects many factors of the information management (searching, extracting, maintaining, uncovering, and viewing information) in addition to the possibilities and efficiency of inference based on semantical information. Requirements for an ontology^{c} language are as follows: 1) a well–defined syntax, 2) a formal semantics, and 3) sufficient and efficient expressive power.

Semantic Web technologies [5, 155, 157] are currently most prominent technologies in the domain of information exchange and reasoning. The goal of Semantic Web is to enable computers to do more useful work and to develop systems that can support trusted interactions over a network. Based on Semantic Web technologies several domain–specific solutions have been developed, such as solutions for healthcare, life sciences, energy, and

sensors. The representation of the syntax of Semantic Web technologies is based on Extensible Markup Language (XML) [160, 161]; though, there are other syntaxes, too. The representation of semantics is based on Resource Description Framework (RDF) [151, 153], RDF Schema [152], and Web Ontology Language (OWL) [154, 158, 159]. There are other options, such as Topic Maps and DARPA Agent Markup Language (DAML) [30, 74].

OWL is a language to represent ontologies^c.[14] OWL is designed to formulate and to reason with knowledge^p about a domain of interest. The conceptual structure of OWL 2 is defined using UML.

There are two variations of OWL–based representation of semantics [159]: OWL 2 direct semantics (OWL 2 DL) and OWL 2 RDF–based semantics. The direct semantics can be utilized in ontologies^c that are OWL 2 DL subset of OWL 2. OWL–based ontologies^c that do not follow OWL 2 DL are set to belong to OWL 2 Full. Direct semantics assigns the meaning for OWL 2 using the style of description logic, and RDF–based semantics is an extension of the semantics for RDF schema (RDF graphs).

OWL 2 Full [159] is undecidable, and it is very complicated to implement a reasoner. Therefore, there are subsets of OWL 2, such as OWL 2 DL, which are designed so that the implementations of reasoners are not overwhelming tasks. OWL 2 specifications comprise three profiles according to application requirements: OWL 2 EL, OWL 2 RL, and OWL 2 QL. The OWL 2 EL profile is designed to be used in the domains that require large ontologies with complex structural descriptions. The OWL 2 QL profile is designed for a standard relational database technology. The OWL 2 RL profile is designed to be used in the application domains that require scalable reasoning and still maintaining as much expressive power as possible.

OWL 2 is based on predicate logic, and it does not yet support modal logics. But there is a W3C Candidate Recommendation of Time Ontology in OWL [156]. In general, OWL 2 specifications assume that information is true, and they do not make any difference between belief^p, justified belief^p, and knowledge^p. As such we see that OWL 2 is the first step towards the presentation of information with different epistemic quality and supporting modal logics; however, there will be several requirements to enhance OWL.

An example of an OWL ontology describing the scenario of Traffic Alert Service (see page 14) is in Appendix *Discussions on Evaluating Epistemic Quality of Beliefs*.

[14] Ontology^c is an explicit and formal specification of a conceptualization. An ontology^c describes formally a domain of discourse. Ontology^p is the study of the nature of existence, which is concerned with identifying the kinds of things that actually exist, and how to describe them.

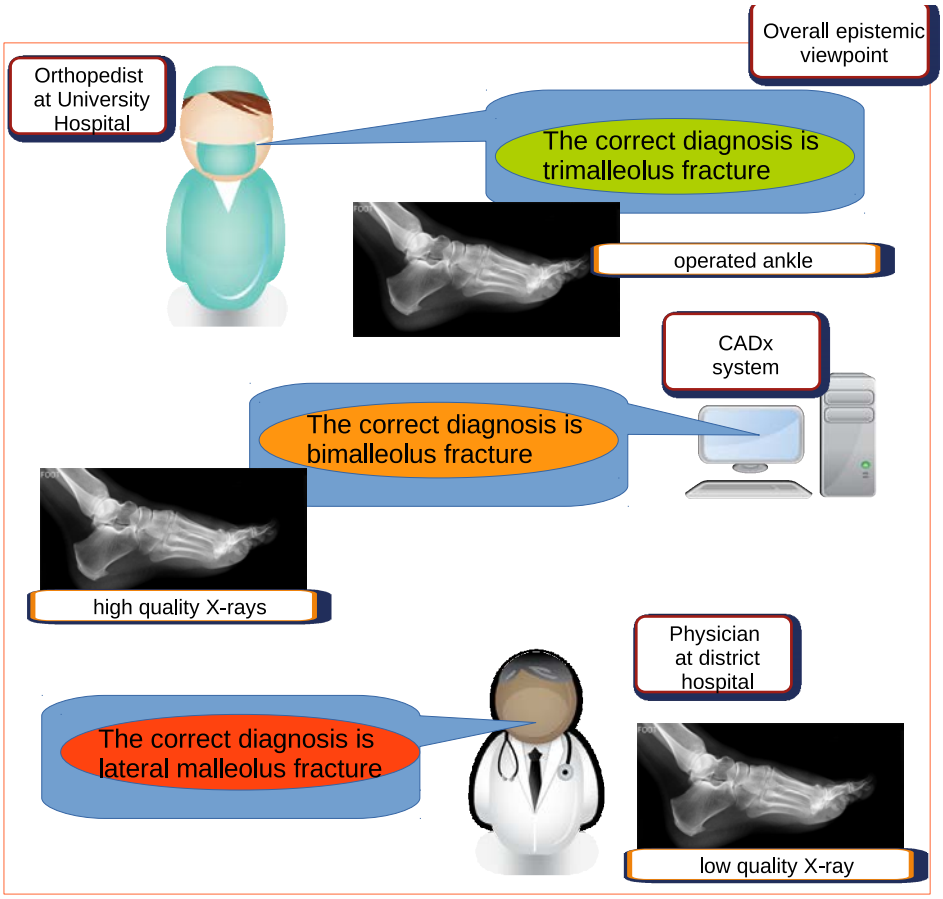

Figure 2.7: Different contexts of propositions.

2.2.2 Knowledge and Justified Belief in Dependable Intelligent Distributed Systems

In this section we discuss the role of knowledge^p, justified belief^p, and belief^p in the context of DIDS. We utilize the scenario *A Tourist having an Accident in a Foreign Country* introduced in Section 2.1.1 to illustrate our approach by analysing the roles of belief^p, justified belief^p, and knowledge^p in each phase to achieve dependability. The scenario illustrates co–operation between ISA$_{bdi}$s[15] and human beings.

In this scenario—actually in real life, too—the propositions stating the

[15]We assume that CADx, THIS, and TMA are implemented using ISA$_{bdi}$s (hybrid model).

diagnosis have the most important role in the process of Matti's medical care. Therefore, we need to evaluate the epistemic quality of information stated by the following propositions:

- *"The correct diagnosis is lateral malleolus fracture."*
- *"The correct diagnosis is bimalleolus fracture."*
- *"The correct diagnosis is trimalleolus fracture."*

And what could be the effect of each epistemic quality of information on the process of Matti's medical care?

When we evaluate the epistemic quality of information stated by the propositions, we have to consider several different viewpoints and contexts, such as (see Figure 2.7):

1. Medical personnel in a small town and a hospital in the small town. The level of professional skills in medicine is not in accord with modern, high–level expertise, and the quality of medical equipment is low–level.
2. Medical personnel at EMA and travellers' emergency medical assistance. The professional skills in medicine and the equipment represent high–level expertise in analysis of potential medical risks in the treatment of patients abroad and the repatriation of patients.
3. CADx and AI. CADx represents a high–level, the state–of–art expert system to analyse X–rays, especially to reveal bone fractures.
4. An orthopedist at Helsinki University Hospital. The level of professional skills in orthopaedic is in accord with modern, highest level expertise.
5. Travel insurance company and compensation for disabilities. The combination of good service and overall cost reduction are the key issues in this viewpoint.
6. Matti's recovery and his future occupation. To avoid Matti's disability and to ensure his possibility to continue his acting career in action movies are the key issues in this viewpoint.
7. An attributor's[16] viewpoint. This represents the viewpoint of an epistemologist who has no role in the scenario itself. This is a kind of an 'objective' viewpoint.

[16] An attributor is a person who evaluates the epistemic quality of information from the external viewpoint on the scenario.

Let us look at the epistemic quality of information stated by the propositions from the viewpoint of Matti's recovery and his future occupation. The most important factor is the correctness of the diagnoses, because otherwise proper medical treatment and full recovery are severely jeopardized. As we assume proper justification and knowledge theories to be forms of reliabilism we concentrate on the reliability^p of the belief^p–forming processes and possible defeaters created by a reliable^p belief^p–forming process.

The proposition *"The correct diagnosis is lateral malleolus fracture."* is the object of the belief^p that Matti, the physician at the district hospital, and the physician at EMA have in phase 1. But is the belief^p justified belief^p?

We can consider that the physician at the district hospital has a justification for his belief^p, as he has been trained to rely that the process which has produced the X–ray image is reliable^p, the quality of the X–ray image is adequate, and his capability to interpret the X–ray image is reliable^p enough. Thus, in his context the reliability^p of the diagnose process is high enough, and there are no known defeaters in this phase (*reliabilism*). The physician ensures Matti that the diagnosis is the correct one; thus, in phase 1 there is also justification for Matti's belief^p because there are not yet any defeaters (*testimonially transferred*). As Alvin Goldman states: *"Surely a belief can sometimes be justified even if additional evidence–gathering would yield a different doxastic attitude."* [60]. However, the physician at EMA has no justification for the belief^p, because based on the former experiences about similar cases she has learned that this kind of diagnoses must be verified by further examinations. This is because the reliability^p of the belief^p–forming processes (e.g. the quality of X–ray image and the physician's medical skills) are not at the required level (*defeater*). The piece of information stated by the proposition *"The correct diagnosis is lateral malleolus fracture."* is not knowledge^p because the proposition is not true.[17]

In phase 2, the defeater (the second diagnosis) is informed to Matti; therefore, the first diagnosis is no longer justified, and Matti formed the new belief^p. The proposition *"The correct diagnosis is bimalleolus fracture."* is the object of the belief^p that the CADx system, Matti, TMA application, and the physician at EMA have in phase 2. The belief^p is also the justified belief^p in phase 2 because

CADx system: The belief^p has been formed using a reliable^p process, and there are no known defeaters (*reliabilism*).

[17] But in this phase the epistemic agents do not have any factors to determine that the diagnosis in not true.

Matti: He has obtained the belief^p from a reliable^p source, and there are no known defeaters (*testimonially transferred*).

TMA application: It has obtained the belief^p from a reliable^p source, and there are no known defeaters (*testimonially transferred*).

The physician at EMA: She has obtained the belief^p from a reliable^p source, and there are no known defeaters (*testimonially transferred*).

However, the piece of information stated by the proposition *"The correct diagnosis is bimalleolus fracture."* is not knowledge^p because the proposition is not true.

In phase 3 there is the defeater—the third diagnosis—that cancels the justification of Matti's, the physician's and TMA's belief^p. The third proposition *"The correct diagnosis is trimalleolus fracture."* is the object of the belief^p that Matti, the TMA application, the physician at EMA, and the orthopedist at the Helsinki University Hospital have in phase 3. The belief^p is also the justified belief^p in phase 3 because

Orthopedist: He has obtained the belief^p from the reliable^p process (the operation of the ankle) and there are no defeaters (*reliabilism*).

Matti: He has obtained the belief^p from the reliable^p source, and there are no defeaters.[18] (*testimonially transferred*).

TMA application: It has obtained the belief^p from the reliable^p source, and there are no defeaters (*testimonially transferred*).

The physician at EMA: She has obtained the belief^p from the reliable^p source, and there are no defeaters (*testimonially transferred*).

The belief^p, the object of which is the proposition *"The correct diagnosis is trimalleolus fracture."*, is knowledge^p. It is true based on perceptions, the reliability^p of which is regarded to be the highest possible (*the correspondes theory of truth*).[19]

However, when having the attributor's viewpoint to the epistemic quality of information stated by the propositions, we have a different result of the analysis, as follows:

[18] Most likely there will not be any defeaters in the future.

[19] However, we can not claim, that this knowledge^p is factual because there is always a possibility that the orthopedist may have made an error; though the possible error may never be revealed because the results based on the knowledge^p are good.

First diagnosis: We can argue that the proposition is the object of merely a belief^p because it is not formed using a reliable^p enough process and there exist defeaters. The epistemic agents in question just are not aware of those issues.

Second diagnosis: We can argue that there is no justification for the belief^p, even though the belief^p has been formed using a reliable^p process, but the reliability^p of which is not high enough when considering the requirement of reliability^p set by the possible consequences of failure. There exists the defeater of which the epistemic agents in question just are not aware.

Third diagnosis: We can argue that there is justification for the belief^p because the belief^p has been formed using the highly reliable^p process.[20] There exist no defeaters. The belief^p is also true, if we accept, for instance, the correspondence theory of truth (see Section 3.5.3) and consider that the correspondence can be verified using a reliable^p perception. Therefore, the third and final proposition *"The correct diagnosis is trimalleolus fracture."* is knowledge^p.

As a final remark of this scenario, we have argued that the role of belief^p, justified belief^p and knowledge^p in DIDS based systems is essential, and therefore, knowledge^p has value over justified belief^p and belief^p, but in some cases those are very difficult to grasp. There are several issues, such as context–awareness and the requirement of factualism.

In DIDS (multi–ISA$_{bdi}$ systems) there is also involved the aspect of social rationality, which can be expressed using the following assumptions [3]:

1. Sincerity: No ISA$_{bdi}$ will attempt to have others believe a proposition that it either knows or believes to be false or a proposition that it wants to be false.
2. Honesty: ISA$_{bdi}$ must act according to their beliefs.
3. Fair Play: ISA$_{bdi}$ must abide by the agreed deals.
4. Sociability: In the case of indifference, ISA$_{bdi}$ must accept others' offers, and deals must always be individually rational.

In principle, these social rationality assumptions would make the development and execution of multi–ISA$_{bdi}$ systems much easier, cheaper, and

[20] Even though the process is very highly reliable, there is always a minor possibility that there is an error in the diagnosis.

cost–effective, but the reality of today's Internet world—which will be the dominant environment for ISA_{bdi}s—does not support these assumptions, at all. Thus, it is increasingly important to develop ISA_{bdi}s that are not just powerful, but also transparent to inspection regarding their knowledge^p, justified beliefs^p, and beliefs^p[21] [22].

2.2.3 Logical Issues of Knowledge, Justified Belief, and Belief

In this section we discuss some features of epistemic logic in order to provide background information on reasoning methods related to belief^p, justification, and knowledge^p.[22]

AI, especially GOFAI, has been heavily influenced by ideas of philosophical logic when trying to solve or solving problems dealing with knowledge^p representations, reasoning, and communications between ISAs [145]. The fundamental theoretical topics, such as epistemic logic, temporal logic, and belief^p revision in addition to the formalization of non–mathematical reasoning are similar to both AI and philosophical logic [145]. Predicate logic and various modal logics[23] play important roles in this co–operation, and AI has affected in many ways recent developments of modal logics.

One of the main differences between AI and philosophical logic is that the latter one does not, in general, deal with implementability or efficiency of reasoning, whereas to the former one these issues are important, as applications of above–mentioned logics are most often the driving force in AI [145]. Examples of the application domains comprise understanding narratives, diagnosis of various failed entities, spatial reasoning, and reasoning about the attitudes of other agents. The ultimate goal has been to formalize common–sense reasoning [145]. But there are still ongoing debates about the suitability of logic to solve problems in AI. For example, it is very difficult to express using logic analogy, space, shape, and uncertainty. And the performance requirements of AI systems are often higher than contemporary logic–based inference systems can fulfil [4].

Modal logic deals with reasoning that involves expressions such as *necessary*, *possibly*, *obligatory*, *permitted*, etc. In the context of ISA_{bdi} the

[21] Of course the same applies to inferring methods.

[22] The logical framework of BDF (see Chapter 6) is outside the scope of this thesis, therefore, we do not discuss in more detail, utilize, or further develop epistemic logic or other modal logics. However, we see that knowing the relevant logical issues is beneficial in understanding the context of this thesis.

[23] The term 'modal logic' seems to have two different scopes: 1) only alethic logic and 2) in addition epistemic, temporal, action, etc. logics.

important logics are in addition to predicate logic the following ones:

1. Epistemic logic, which studies reasoning in the domain of belief^p, justification, and knowledge^p.

2. Temporal logic, which studies reasoning in the domain of time.

3. Logic of actions, which studies reasoning in the domain of actions.

In this thesis we concentrate on epistemic logic.

G.H. von Wright's article *An Essay in Modal Logic, 1951* [150] can be considered as a starting point of the formal study of epistemic logic as it exists today [115]. Jaakko Hintikka demonstrated in his book Knowledge and Belief [70] that the modal approach to single–agent epistemic attitudes is beneficial. He explicated epistemic attitudes using a model of theoretic relation over possible worlds [145].

Epistemic logic focuses mainly on propositional knowledge^p and beliefs^p. Justification has not yet raised any significant interest. The language of epistemic logic is based on the language of propositional logic that is enhanced with knowledge^p and belief^p operators. K_1A means that agent 1 knows A and B_2A means that agent 2 believes A for an arbitrary proposition A. One of the differences between the logic of knowledge^p and the logic of belief^p is that the logic of knowledge^p includes the schema (T) $K_iA \rightarrow A$ stating that knowledge^p must be true, but the logic of belief^p does not require a proposition to be true—a doxastic agent may have false beliefs^p.

Time plays a fundamental role in the actions of ISA$_{bdi}$; therefore, temporal logics are also important in the context of ISA$_{bdi}$. There are two approaches to temporal logics, of which one is based on modal logic and the other one is based on predicate logic. The approach based on modal logic was introduced by Arthur Prior in the 1960s [61]. The languages comprise the following four operators:

- **G** : "It will always be the case that ... "
- **F** : "It will at some time be the case that ... "
- **H** : "It has always been the case that ... "
- **P** : "It has at some time been the case that ... ".

Michael Bratman's philosophical analysis of the notion of intention is one of the starting points of the logic of action, especially in AI and computer science. Practical reasoning is reasoning directed toward actions, thus

the process of figuring out what to do [165]. Agents' intentions play the important role when selecting actions that are desired and when committing to the selected actions. Rao and Georgeff [131] formalized the belief–desire–intention model using the branching–time temporal logic **CTL**, on top of which they introduced modal operators for *belief, goal (nowadays desire), and intention* as well as operators for the results of actions *succeeded(a) and failure(a)*. The formal semantics is based on the Kripke models with accessibility relations between worlds for each modal operator.

The semantics of modal logics is defined by using possible worlds semantics, where a set W of possible worlds w is defined. There is a truth value assigned to each propositional variable in the specified language for each of the possible worlds w in W. For example, the truth value of an atomic proposition p at world w, $w \in W$, given by the valuation Γ can be expressed as $\Gamma(p,w)$. Then the truth value of a complex proposition of modal logic for a given valuation Γ and a world $w \in W$ can be specified as for example: $\Gamma(\Box A, w) = T$ *iff for every world* $w' \in W, \Gamma(A,w') = T$.[24] Thus $\Box A$ is true at a world w exactly when A is true in all possible worlds [50].

Next we discuss some requirements for modal logics in the context of ISA_{bdi}. These requirements are only a highlight of the issues, which need to be addressed. In addition to normal philosophical requirements of logic (soundness and completeness) there are requirements such as efficiency, practical reasoning, common–sense reasoning, and a philosophical aspect how to actually characterize the properties of ISA_{bdi} in terms of formulae of modal logics.

Efficiency requirements are usually set by the performance requirements of services that ISA_{bdi} is designed to provide. Especially, if ISA_{bdi} operates in a real–time environment, the implementation of belief^c, desire^c, and intention^c databases—for example, representations of beliefs^c—must be efficient in addition to efficient reasoning processes. ISA_{bdi} cannot deliberate indefinitely. However, we do not discuss in this thesis how these requirements could be achieved.

Practical reasoning distinguishes from theoretical reasoning in that theoretical reasoning is directed towards beliefs^p, but practical reasoning directed towards actions [165]. Practical reasoning seems to be a two–phase activity: at first, ISA_{bdi} deduces what states of affairs it wants to achieve, and then ISA_{bdi} performs means–ends reasoning about how to achieve the selected state of affairs. This implies that there is a need for a kind of hybrid logics.

[24] $\Box$ == *'it is necessary that'*

Currently common understanding is that usable AI and thus ISA_{bdi} exhibiting human–like behaviour requires common–sense knowledge [87]. Common–sense knowledge is difficult to define, but usually it is seen as a collection of simple facts about everyday life such as "dogs bark" and "cats meow". In fact, common–sense reasoning also requires a combination of different modal logics. In general, common–sense reasoning does not require that all the possibilities to accept a proof must be satisfied, and very improbable possibilities can be neglected. This refers to the so–called frame problem. Mostly, this is due to the required efficiency of the reasoning as all the logically possible options cannot be processed in a required response time or using available computing resources.

When we try to characterize properties of ISA_{bdi} in terms of formulae of modal logics, we deal, for example, with the problem of what is the right formula to characterize the relationship between intentions^c and beliefs^c [165]. For example, how to build up a formula that characterizes "*If* $\boldsymbol{i}$ *intends* φ*, then* $\boldsymbol{i}$ *believes* φ *is possible.*"? Is this the correct one: *(Int i* φ*)* $\Rightarrow$ *(Bel i* $\boldsymbol{E}\Diamond\varphi$*)*[25]?

Epistemic Logic

A modal approach to epistemic logic is in principle simple: systems of modal logic are provided with epistemic interpretation, and main technical results about epistemic logic can be obtained almost automatically. To interpret modal logic epistemically, an epistemic agent reads modal formulas as epistemic statements that express the epistemic agent's attitude towards certain sentences. In addition, the epistemic agent has a new interpretation of the semantics for modal logic [34]. Some features of the logical behaviour of epistemic concepts are quite obvious. For example, claiming to know p and q implies to know q[26], and it cannot be coherent to assert "p *but the epistemic agent does not believe (know)* p" [115]. Furthermore, if the epistemic agent knows p then p must be the case.

However, describing actual knowledge^p is a very difficult task, as actual knowledge^p does not always seem to obey any logic [34]. Therefore, idealizations are required regarding the reasoning capabilities of epistemic agents. This raises a question of the correct level of idealization [34]: An idealized model should correspond the intended environment of an epistemic agent exactly enough in order the epistemic agent to operate according to its service requirements. But, on the other hand, the idealized model should enable the resulting epistemic logic to be weak enough for the (AI–based)

[25] $\mathbf{E}$ is an existential path quantifier, that is $\mathbf{E}\varphi$ is true on some path

[26] $(p \wedge q \rightarrow q)$

epistemic agent, for example, by not requiring it to be a very powerful reasoner, which knows all the logical consequences of it's knowledge, including all logical truths (so–called logical omniscience problem). To solve a part of this problem we need to somehow categorize knowledge^p (and belief^{p}s) into classes that epistemic agents, like ISA$_{bdi}$, can really know and classes that epistemic agents should know, if they had enough resources to do so.

There are several different versions of the epistemic logics:

- Logic of knowledge for single agent reasoning
- Logic of knowledge for multi–agent reasoning
- Logic of common knowledge for multi–agent reasoning
- Logic of justification
- Logics for multi–modal contexts
- Logic of beliefs for agent reasoning
- Logic for knowledge and belief representations.

In this thesis we discuss mainly the first two items and very briefly the third and fourth ones.

Epistemic Logic of Single Agent

The language of epistemic logic comprises the language of propositional logic added with the following unary epistemic and doxastic operators:

- $K_i A$ meaning that agent i knows A and
- $B_i A$ meaning that agent i believes A.

For example, the meaning of the formula $\neg \mathrm{K}_i \neg \mathrm{A}$ is that agent i considers A possible [123].

The language of epistemic logic is defined as follows:

Definition. *Let L be a non–empty, countable set of atomic formulae of the propositional logic and i be an agent. The sentences of L^K are defined inductively as follows:*

1. $L \subseteq L^K$

2. *If* $A \in L^K$ *then* $\neg A \in L^K$
3. *If* $A \in L^K$ *and* $B \in L^K$ *then* $(A \wedge B) \in L^K$
4. *If* $A \in L^K$ *and* $B \in L^K$ *then* $(A \vee B) \in L^K$
5. *If* $A \in L^K$ *and* $B \in L^K$ *then* $(A \rightarrow B) \in L^K$
6. *If* $A \in L^K$ *and then* $K_i A \in L^K$.

The axioms of epistemic logic are specified as follows [34, 115, 123]:

Definition. *K is the epistemic logic specified by the following axioms and rules on inference:*

(PC)*: All tautologies of the propositional logic*

(K)*:* $K_i A \wedge K_i(A \rightarrow B) \rightarrow K_i B$ *(Distribution axiom)*

(MP)*: From A and* $A \rightarrow B$ *to infer B (Modus Ponens)*

(RN)*: From A to infer* $K_i A$ *(Rule of Necessitation)*

We can obtain stronger logics by adding to the logic K principles that express other desirable properties of knowledge^p. For example, the following common schemes can be added [115]:

(T): $K_i A \rightarrow A$ *(Knowledge axiom)*

(D): $K_i A \rightarrow \neg K_i \neg A$ *(Consistency axiom)*

(4): $K_i A \rightarrow K_i K_i A$ *(Positive introspection axiom)*

(5): $\neg K_i A \rightarrow K_i \neg K_i A$ *(Negative introspection axiom)*

(.2): $\neg K_i \neg K_i A \rightarrow K_i \neg K_i \neg A$

(.3): $K_i(K_i A \rightarrow K_i B) \vee K_i(K_i B \rightarrow K_i A)$

(.4): $A \rightarrow (\neg K_i \neg K_i A \rightarrow K_i A)$

Knowledge^p axiom (T) states that knowledge^p must be true. Systems containing the schema (T) are called the logic of knowledge^p, and if the schema (T) is dropped then systems are called the logic of belief^p. In the context of ISA_{bdi}, especially when ISA_{bdi} perceives its external world, it is not straightforward to detect when a proposition is true or false (see Section 3.3). However, care should be taken not to collapse knowledge^p and belief^p in the combined systems [115].

Consistency axiom (D) demands that an agent must be consistent in its knowledge^p; hence, it cannot know both a formula and its negation. In the context of ISA$_{bdi}$, especially in the cases where ISA$_{bdi}$ has a huge knowledge database, the consistency requirement may create problems to efficiency requirements.

The positive introspection axiom (4) states that the agent actually knows what it knows, and the negative introspection axiom (5) states that the agent actually knows what it does not know. In context of ISA$_{bdi}$ these axioms may lead to the regression problem and make the implementation of belief^c databases of ISA$_{bdi}$ complicated by requiring metadata of metadata of ... of knowing that knows.

By selecting separate sets of the above axioms different systems of axioms can be established, and each system of axioms has a different level of modal strength. The axiom (K) establishes the system **K**, but this system is too weak for epistemic reasoning. Hence, the system **T**, which comprises the axioms (K) and (T) as valid axioms, is the weakest one used in epistemic reasoning.

In the context of ISA$_{bdi}$ it is not quite clear what these requirements mean on implementations of ISA$_{bdi}$ as for example, they can easily create a regression problem: ISA$_{bdi}$ knows that it knows that it knows that ... or demand a powerful reasoner that cannot properly be implemented using today's software solutions.

Epistemic Logic of Multiple Agents

Nowadays an increasing number of systems based on ISA are multi–agent systems comprising several—most often different kinds of—epistemic agents. Therefore, there is a need for multi–agent epistemic logic systems. One way to achieve this is to augment a single agent epistemic logic to a group of agents. Hence, there are two primary differences compared to single agent epistemic logics: the number of agents and the number of accessibility relations.

Let us suppose that there is a group **G** consisting of n agents. The language of single–agent epistemic logic [34, 115, 123] is augmented by n knowledge operators K_1, K_2,..., K_n. Now, it is possible to state that *"agent 1 knows that agent 2 does not know that agent 1 knows that p"*: $K_1 \neg K_2 K_1 p$. The model **M** for a single agent epistemic system is augmented by n accessibility relations R_1, R_2,..., R_n. The structure of the model **M** is the following one: $\mathbf{M} = (W, R_1, R_2, \ldots, R_n, V)$. And the satisfaction

relation is as follows:

M,w $\models$ $K_i A$ *iff for every* $v \in$ *W: if* $wR_i v$ *then M,v* $\models$ *A.*

There is a special case regarding a group of agents to know facts, and that is the case when all the agents in the group know simultaneously the very same fact *that A*. This is called common knowledge, and it can be defined as follows [115]: $K_1 A \wedge K_2 A \wedge ... \wedge K_n A$. Common knowledge can be semantically interpreted using the model for multi–agent systems and augmenting it by a special accessibility relation R^c, where $R^c = (R_1 \cup R_2 \cup ... \cup R_n)$. And the model is as follows: $\mathbf{M} = (W, R_1, R_2, ..., R_n, R^c, V)$. The satisfaction relation is *M,w* $\models$ *CA iff* $wR^c v$ *implies M,v* $\models$ *A.*

Epistemic Logic of Justification

The epistemic logic of justification has not yet achieved any significant popularity among AI researchers and logicians. It has not been common to study epistemic justification using logical principles. One exception is the closure principle, which has been discussed in the evaluation of different justification theories. According to conjunction closure the set of propositions that one has justification for believing is closed under the operation of taking conjunction, that is from $\mathbf{JP} \wedge \mathbf{JQ}$ to infer $\mathbf{J(P \wedge Q)}$ [138].[27] This principle and its (in)validity has been discussed to some extent in the contexts of the lottery and preface paradoxes. However, the logic of epistemic justification is still in its infancy. In his article *The logic of epistemic justification* Martin Smith tries to import into debates the nature of epistemic justification according to which logical principles can provide valuable resources for evaluating different theories of justification [138]. Also, Sven Rosenkranz discusses the logical principles of justification in his paper *The Structure of Justification* [119].

Risk minimisation theories claim that *"one has justification for believing a proposition P just in case it would be unlikely, given one's evidence, for P to be false"* [138]. Risk minimisation theories (we can count reliabilism to be one of them) require that there is some probability threshold beyond which a proposition would not be false, and therefore, believing the proposition is justified. Thus, there is a probability function Pr such that E $\Rightarrow$ P iff $\Pr(P|E) > t$, for some t close to but less than 1. Martin Smith states in his article *The logic of epistemic justification* [138] *"A probability function is nothing more than a function mapping propositions to numbers in a way*

[27] J is a modal operator which states that one has justification for believing.

that meets certain constraints. The domain of a probability function is a set of propositions F that is closed under negation and disjunction and includes a maximal proposition entailed by all others in the set."

Possible worlds are often used to model propositions for computing probability functions:

1. $\Pr(W) = 1$ (W = a set of possible worlds)
2. $\Pr(P) \geq 0$
3. If P and Q are inconsistent then $\Pr(P \vee Q) = \Pr(P) + \Pr(Q)$.

Conditional probability is defined by the formula [138]:

$$\Pr(P|E) = \Pr(P \wedge E)/\Pr(E) \text{ if } \Pr(E) > 0 \text{ and is undefined otherwise.}$$

Martin Smith argues that *risk minimisation* theories have problems with conjunction closure [138]. The basic reason is the following one: Let us suppose that the justification threshold is 0.9, and

1. The probability of proposition P1 is 0.95, therefore it is justified;
2. The probability of proposition P2 is 0.95, therefore it is justified;
3. The probability of proposition P3 is 0.95, therefore it is justified;
4. The probability of conjunction of propositions P1 $\wedge$ P2 is 0.9025, therefore it is justified;
5. The probability of conjunction of propositions P1 $\wedge$ P3 is 0.9025, therefore it is justified;
6. The probability of conjunction of propositions P2 $\wedge$ P3 is 0.9025, therefore it is justified; but
7. The probability of conjunction of propositions P1 $\wedge$ P2 $\wedge$ P3 is 0.8573, therefore it is not justified.

This is not according to conjunction closure [138]. But if the threshold is 1 (infallibilist theory of justification), then conjunction closure is assured. But infallibilism is the way to scepticism. Martin Smith argues that the risk minimisation theories invalidates also the following ones [138]:

1. Cumulative transitivity: *((E $\Rightarrow$ P), (E $\wedge$ P) $\Rightarrow$ Q): (E $\Rightarrow$ Q)*
2. Monotonicity: *(E $\Rightarrow$ P): ((E $\wedge$ Q) $\Rightarrow$ P)*

3. Amalgamation: *(E ⇒ P), (F ⇒ P): ((E ∨ F) ⇒ P)*

The question is that is it mandatory for the validity of justification theory that conjunction closure and other ones must be assured.

Martin Smith proposes the following alternative theory to the risk minimisation: *"One has justification for believing a proposition P just in case it would be **abnormal**, given one's evidence, for P to be false."* [138]. He argues the above logical principle are invalidated by this theory. However, it is not quite clear what the term *abnormal* means.

Sven Rosenkranz argues that the structural account of justification can be expressed using the following five principles [119]:

E: JP → ¬K¬KP

T$_K$: KP → P

RN$_K$: If ⊢ P, then ⊢ KP

RM$_K$: If ⊢ P → Q, then ⊢ KP → KQ

Lum: ¬K¬KP → K¬K¬KP

Sven Rosenkranz says that the logic of justification based on the above principles involves idealisations. He also argues that the logic of justification cannot be a normal modal logic. The dominant factor is that whether the logic for justification agglomerates over conjunction or not [119].[28]

Sven Rosenkranz also discusses the following principles for justification [119]:

- From **JP** to infer **JJP**.
- From ¬**JP** to infer **J**¬**JP**.
- From **JJP** to infer **JP**.
- From **J**¬**JP** to infer ¬**JP**

Because of the scarce interest in the logic of justification it is not yet at the demanded level, and further studies are required.

[28] For the proof, see the article *The Structure of Justification* [119].

Summary of Logical Issues

Epistemic logic can be categorized into following types: the logic of belief^p, the logic of justification, and the logic of knowledge^p. The first one and the last one have so far gained the most interest among logicians and AI people. The logic of justification is still in its infancy requiring further studies.

In addition to normal philosophical requirements of logic there are requirements such as efficiency, practical reasoning, common–sense reasoning, and a philosophical aspect on how to actually characterize the properties of ISA$_{bdi}$ in the terms of formulas of modal logics. There have been several approaches to solve the problems created by the above requirements, but there is still a lot of research to do. Incomplete information, different degrees of justifications (reliability^p of belief^p–forming process), concurrency, and continuous change are still major problems in epistemic logics in the context of ISA$_{bdi}$. In the context of AI logic quite often requires to undertake the task of formalizing large examples involving non–trivial common–sense reasoning.

The *frame problem* has caused a lot of work on reasoning in AI. The problem arises because it should be necessary for a rational agent to know thoroughly the whole state change: not only the properties of the world that change as the result of an action but also the properties that do not change when the action is executed. Another aspect is, how could ISA$_{bdi}$ limit the scope of the proposition it needs to reconsider in the context of its actions, especially in the cases where ISA$_{bdi}$ has enormous knowledge^p bases to examine?

When we try to implement an ISA$_{bdi}$, which should exhibit human–like behaviour, it is not enough to utilize only epistemic logics because a rational behaviour is most often the result of reasoning concurrently using many different modal logics. Therefore, we need a kind of hybrid logic that combines alethic logic, epistemic logic, temporal logic, and logic of action. As an example of the hybrid approach is Michael Wooldridge's *LORA – Logic Of Rational Agents* [165]. *LORA* extends full first–order branching time temporal logic with the addition of modalities for referring to the beliefs, desires, and intentions of agents, and with a dynamic logic for reasoning about actions. The semantics of *LORA* is very complicated, and it requires a lot of study in order to properly implement it. For more information about *LORA*, see [165].

Some of the problems of modal logics could be overcome using solutions based on connectionism. Therefore, an approach comprising different architectures of ISAs may provide a most proper solution for many contexts and environment of IDS.

Chapter 3

Six Concepts

3.1 Introduction to the Six Concepts

In this chapter we discuss six concepts that we consider to be important in the context of the dependability of IDS. We analyse various epistemological theories and their applicability to the environment of ISA_{bdi} and DIDS. When analysing epistemological theories and their applicability we need to have a new viewpoint in addition to the viewpoint of traditional epistemology. In traditional epistemology we study, define and analyse epistemological theories in the context of an existing system (human brain and natural languages, mainly English) in order to define theories that best address the epistemological problems. In the contexts of ISA_{bdi} and DIDS we need to take into account a viewpoint how to implement theories for new AI systems. This viewpoint comprises several requirements such as how to specify and code programs (programming languages, data presentation languages, ontology languages, data structures, modal logics, ontologies, etc.) and runtime requirements (memory, performance, dependability, infrastructure requirements, etc.). The actual importance of each of these requirements is mostly dependent on applications. In this thesis we do not discuss any specific real applications; therefore, we have a generic approach to these requirements and we do not discuss them in detail. Based on the analysis we propose new definitions of justified belief^p and knowledge^p so that they can be better applied to the environment of ISA_{bdi} and DIDS.

Epistemology is an open–ended study of human knowledge and justified beliefs. Man has discussed knowledge and justified beliefs over two thousand years considering that those terms are attributable only to human beings. However, the progress in computer science and related disciplines during the last 60 years has brought a new dimension to the discussion.

Alan Turing brought the topic on the scene in his 1950 paper *Computer Machinery and Intelligence* [146], where he issued a question: Can machines think? The next fundamental article was John McCarthy's *Ascribing Mental Qualities to Machines* [89] (published 1979), where McCarthy introduced ideas to ascribe certain beliefs, knowledge, free will, intentions, consciousness, abilities or wants to a computer program, when such an ascription expresses the same information in the contexts of a machine and a person. Thus, anthropomorphism had entered into computer science, and we discuss this issue in more detail below.

Significant progress has taken place since those articles, and for example, at the 2014 Turing Test the winning bot—chatting robot—fooled over 30 percent of the judges to think that they were communicating with a human being[1], IBM's Watson won Jeopardy! Challenge, Google's AlphaGo won the world's best GO player (Lee Sedol), and AlphaGo Zero has learned by itself the best moves of the game GO by playing millions of games against itself. Hence, we can claim that intelligent software entities have shown human–like behaviour; therefore, indicating a possibility to have beliefs^p, justified beliefs^p, and knowledge^p. Epistemology and intelligent software entities have been discussed in more detail from different viewpoints, for example, in [4, 6, 21, 106, 135, 134, 163].

John R. Searle raised severe dispute about the capabilities of a computer system to be a mind; thus, to understand, to have intentions, etc. In his Chinese Room argument the main theses were as follows: (1) Intentionality[2] in human beings is created by causal features of the brain and (2) instantiating a computer program is never by itself a sufficient condition of intentionality [128]. One of the main targets was the view that formal computations on symbols could produce thought; in other words, there is no way to attach any meaning to the formal symbols because syntax and internal connections are insufficient for semantics [27]. There are several replies to John R. Searle's claims comprising, for example, so–called virtual mind reply, systems reply, robot reply, brain simulator reply, and intuition reply [27].

We see that the main weaknesses of John R. Searle's arguments are as follows: 1) Terms, such as understand, intention, and belief^p are not unambiguously explicated, well enough. Therefore, there are several grey

[1] Though there are opinions that the Turing Test is not a good test to determine whether a machine can think or not.

[2] *"Intentionality is the power of minds and mental states to be about, to present, or to stand for things, properties, and states of affairs. To say of an individual's mental states that they have intentionality is to say that there are mental representations or that they have contents"* [75].

areas within which it seems to be very difficult to achieve any consensus. 2) John R. Searle seems to assume that human beings have some kind of higher level biological (metaphysical) entity that explains the 'superior' features over artificial entities.[3] Therefore, intentionality (including beliefs^p) is something that seems to be possible only for human beings. 3) The Chinese Room argument is outdated in the context of AI, because AI researchers are more interested in and discussing the term *superintelligence*. Superintelligence refers to artificial entities that greatly outperform the best current human minds in most general cognitive domains. As Nick Bostrom in his book *Superintelligence: Paths, Dangers, Strategies* states that AI needs not resemble a human mind much, and AI will have very different cognitive architectures than biological intelligences [21]. We have discussed the Chinese Room argument in more detail in the Appendix *Is it Time to Get Out of the Chinese Room*.

Our approach to the issues raised by anthropomorphism can be summarized in the following claims:

- Externalist attitude: The states of a physical entity get their content through causal connections to the external reality they represent. This is not limited only to human beings.[4]
- A computer might have propositional attitudes if it has the right causal connections to the world.[5]
- The syntactically specifiable objects over which computations are defined can possess semantics; it is just that the semantics is not involved in the specifications [117]. In the context of ISA$_{bdi}$ we can have semantics involved using metadata about information explaining causal connections to the world and the causal connections themselves.[6]
- Programming is precisely what could give something a mind.[7]

However, anthropomorphism still seems to be a difficult issue, especially from a humanistic point of view, as Sir Anthony Kenny discussed in his first Georg Henrik von Wright lecture at the University of Helsinki [77], and Peter Hacker tried to prove in his talk at the Wheatly Forum [64].

[3] We are not aware of any evidence that would prove human beings to have such an entity.

[4] Corroborated by Fred Dretske, Hilary Putnam, and Jerry Fodor

[5] Corroborated by Jerry Fodor

[6] Corroborated by Georges Rey

[7] Corroborated by Daniel Dennett

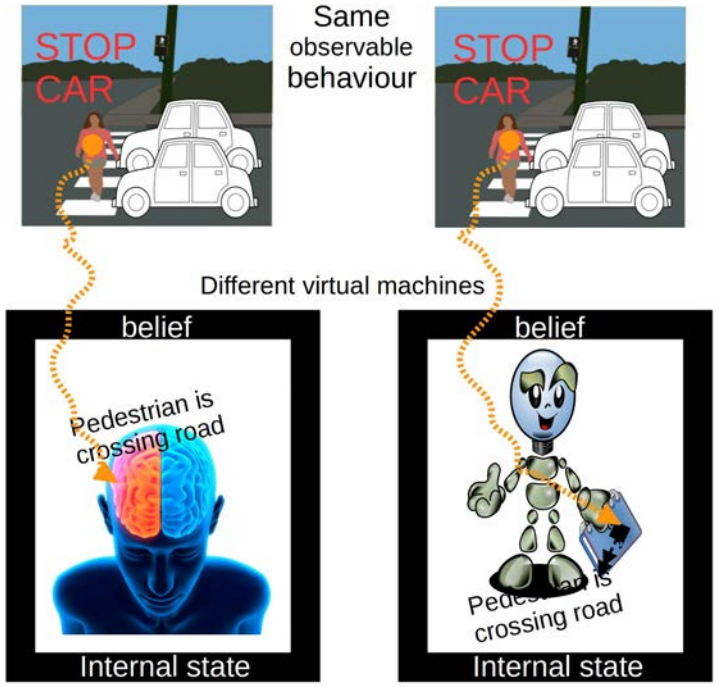

Figure 3.1: Human belief and ISA belief.

We have a black box approach to terms that can be considered to be anthropomorphic. The following example illustrates our thoughts (Figure 3.1):

Let us suppose that two cars are approaching side by side a crossroad, where a pedestrian is just crossing the road so that the cars must stop to give way to the pedestrian. One car is driven by a man and another car is driven by an ISA. Let us also suppose that the status of information and information processing of both the man and the ISA regarding traffic laws is equal. The ISA has also similar kinds of concepts of the pedestrian, crossing and road as human beings have. Now, the man sees via his reliable visual system that the pedestrian is crossing the road and infers based on this perception a proposition *"A pedestrian is crossing the road."* and forms a belief^p based on it. Therefore, he must stop his car in order to avoid an accident. Similarly, the ISA perceives via its equally reliable video, radar and shape recognition system that the pedestrian is crossing the road and deduces based on this perception a proposition *"A pedestrian is crossing the road."* and forms a belief^p based on it. Therefore, it must stop the car in order to avoid an accident.

Now, we can argue that in the case of the man the proposition *"A pedestrian is crossing the road."* is the object of the man's belief^p (see definitions below in Section 3.4). Now, what can we say in the case of the ISA? Clearly, the proposition *"A pedestrian is crossing the road."* is what the ISA considers to be the case, and there is very little unreliability^p of the matter. In addition, the ISA is in a state of having a representation of the proposition stored, and the representation is created by actual and causal relations to sensory stimulations. If we want to explain this situation of the ISA to another person then the concept of belief^p is the best one to explicate it.

What are the differences between these cases? The processes leading to the representations are different, but their outcomes are similar. The observable behaviour in both of these cases is similar. The internal representations of the propositions are quite different. In the case of the man we do not actually know what is the exact (neurobiological) representation of the belief^p, the proposition and its supporting information, because the related sciences (cognitive neurobiology, psychology, etc.) are not yet advanced enough. But in the case of the ISA we do know it. Hence, it seems to be that in this case humanism concerns something that we do not fully understand yet. Therefore, in the case of human beings there might be something—but we don't know exactly what—that deserves the humanistic attitudes towards terms like belief^p, intention^p, knowledge^p, etc. But the meaning of the term *belief*^{p} is the same in both cases; both the man and the ISA are in the state of belief^p, the object of which is the proposition *"A pedestrian is crossing the road."*.

Jerry Fodor argued that human beings are semantic engines with a language of thought [41]. Aaron Sloman and his group at the University of Birmingham have developed a concept of information–processing virtual machine (hereinafter VM) and virtual machine functionalism (hereinafter VMF) [136, 137]. We can use virtual machines to represent the black boxes in Figure 3.1. According to this concept the human mind is one kind of virtual machine, which is operated by a human body. And ISA is another kind of virtual machine, which is operated by a computer. The basic idea of functionalism is that the essence of a mental state is not to be found in the biology of the brain but rather in the role that it plays in one's mind and in the causal relations that it bears to stimuli [6]. Thus, functionalism claims that mental states are not only physical states, but also functions or operations of those physical states. Hence, mind could be implemented in any physical system (natural or artificial), which is capable of supporting the required computation and the functioning of the system including its

actions. Another similar approach is the computational theory of mind (hereinafter CTM), which was proposed by Hilary Putman and further enhanced by Jerry Fodor [112, 116]. But instead of speaking about occurrent propositional attitude states VMF concentrates on the architecture of the mind and the states and functionalities enabled by the components of the virtual machines.

Every VM has an architecture that provides tools to operate on information. The architecture comprises forms of representation, algorithms, concurrently active sub–systems, connections between sub–systems, and causal interaction between sub–systems, etc. The human mind has one kind of architecture and ISA has another kind of architecture. VMF allows multiple, concurrently active, interacting mental states, and an individual can have many mental sub–states at any time. Each sub–state is defined by its causal relationship to other sub–states and its environment [136].

VM schema and VMF provide us with a good, acceptable foundation to discuss epistemic terms in the context of ISA_{bdi}.

Another popular approach to capturing and understanding the mechanism and processes of cognition is to build models using networks of simple, neuron–like processing elements, each of which performs simple numerical computations as already discussed in Section 2.2.1. In this approach a representation is often a pattern of activations over a set of processing units in the model [144]. But this form of the representation of information is much more difficult to analyse and evaluate than the representations based on explicit symbols; hence, it is not suitable for the objectives of this thesis.

Intuition^p plays a significant role in epistemology [111]; for example, when evaluating the appropriateness of different definitions of knowledge^p, justified belief^p, and belief^p. However, the exact nature of intuitions^p has not been precisely explicated, and principled taxonomy of the various kinds of intuitions^p has not been established [111]. Therefore, we can claim that there is no significant role of intuition^p in the context of ISA_{bdi}. At first glance this seems to be in contradiction with our claim above: ISA_{bdi} can have knowledge^p, justified beliefs^p, and beliefs^p. But this is not the case. Intuition^p has a role in the evaluation of theories and definitions, but not, when evaluating whether the information stated by a proposition is knowledge^p and/or justified belief^p based on an appropriate definition. Therefore, ISA_{bdi} is not required to have intuition^p.

We can summarize our basic thoughts on the philosophy of the mind and anthropomorphism in the context of ISA_{bdi} as follows: we support virtual machine functionalism, which implies that same mental states can be realized using quite different methods and various physical systems, such

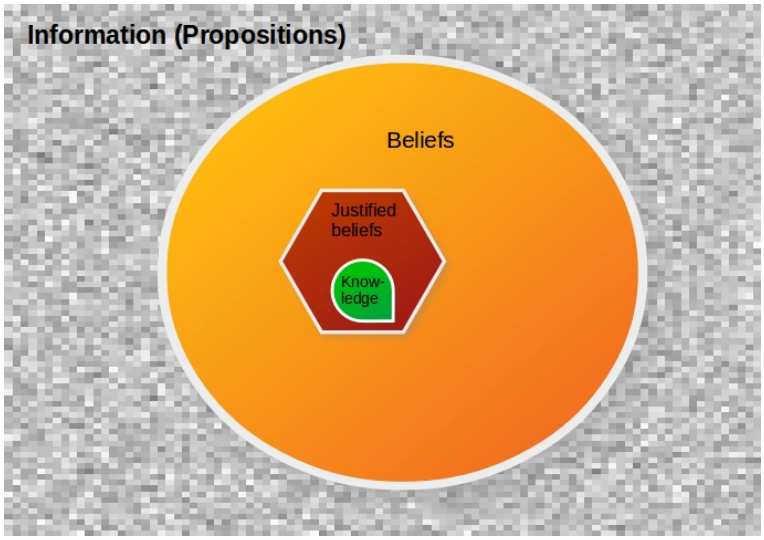

Figure 3.2: Classification of information.

as a human brain, a dolphin brain, and an artificial entity. Therefore, we argue that we can use anthropomorphic terms, such as belief^p, justified belief^p, knowledge^p, and intention^p in the context of artificial entities, such as ISA$_{bdi}$. In addition, we support speech–act theory, which emphasizes that the key unit of linguistic meaning is not an abstract sentence but an utterance as a concrete act carried out by people and ISA$_{bdi}$s [6].

We have the following classification of information as illustrated in Figure 3.2. There are many different meanings for the term *information* depending on the context of its usage, for example, *information is that which informs* [33, 40]. A general definition of information as a semantic content that comprises both data and meaning is the following one [40]:

***I** is an instance of information, understood as objective semantic content, if and only if:*

1. ***I** consists of n data (d), for n ≥ 1;*
2. *The data are well–formed;* and
3. *The well–format data is meaningful.*

In this thesis we define *information* to be any data that is syntactically well–formed and has a meaning—that is, it involves semantic—to an epistemic agent, who holds information.

In philosophy proposition seems to be a difficult concept, and even the existence of proposition is questionable as discussed in the article written by Matthew McGrath and Devin Frank [90]. The term *proposition* has many different explications. A proposition can be used as the primary bearers of truth–value, the object of beliefs, the referents of that–clauses, and the meanings of sentences [90]. In this thesis we use the term *proposition* to be shareable objects of attitudes and the primary bearers of truth–value. We also see propositions to be a type of objective semantic contents. Semantic information has an important role in communications, where an important type of semantic information is factual information. It tells the informee something about something else. Factual information has a declarative nature [40].

Belief^p is the attitude of an epistemic agent to a proposition that the epistemic agent considers to be true. Justified belief^p is belief^p for which the epistemic agent has a justification for it to be true. Knowledge^p is justified belief^p that is true.[8] [9]

Next we discuss in more detail epistemic value, the theories of truth, belief^p, justified belief^p, knowledge^p, and trust. We start with a discussion of epistemic value in order to point out issues why knowledge^p is more valuable than plain (true) belief^p. We continue the discussion with the topic of truth, because truth is a kind of fundamental concept in the way that epistemic agents aim at it or at least should aim at it. On the other hand, it is difficult to comprehend the actual meaning of truth in the context of ISA$_{bdi}$.

3.2 Epistemic Value

One of the aims of epistemology is to understand the value of knowledge^p. Is knowledge^p valuable, and if it is valuable then why? The same question can be raised in the context of ISA$_{bdi}$, as well. We argue that knowledge^p is more valuable than justified belief^p also in the context of ISA$_{bdi}$, but not exactly in the same sense as discussed in the Meno problem written by Plato. At first, we need to clarify what we mean with the term *epistemic value*, as its explication varies across contexts. There are at least two different ways to discuss epistemic value [108, 109]: The first way to express

[8] There are also other requirements for belief^p to be knowledge^p as discussed in Section 3.6.

[9] As we support fallibilism, we are of the opinion that there is a minor possibility that a belief^p which we reason to be knowledge is not always true. For example, science is considered to aim at and create knowledge, but history has shown that many times scientific knowledge is proved later to be untrue.

that something is of epistemic value is to say that it is something which is epistemic (for example knowledge^p) and which is valuable. The second way to express that something is of epistemic value is to attribute a particular kind of value to it (a kind of value which is epistemic). There is a difference between epistemic value and the value of epistemic. It should be noted that something epistemic can also have non–epistemic values, such as an aesthetic value. In the context of ISA$_{bdi}$ we are primarily interested in the first interpretation, as it expresses how epistemic entities are utilized by ISA$_{bdi}$ when it operates to provide services. For example, the proposition *"Snow is white."* is seen as an epistemic entity having epistemic quality of belief^p, justified belief^p, or knowledge^p, and it has a value—various degrees based on its epistemic quality—in the process of reasoning about the next action to carry out.

In recent decades a lot of attention has been devoted to the question *"Why does knowledge matter?"*; especially, why knowledge^p *that p* is more valuable than mere true belief^p *that p*, or actually, why knowledge^p *that p* cannot be more epistemically valuable than mere true belief^p *that p*. At least, we can argue that truth in one's belief^p is minimally valuable in the sense, that all other things being equal, true beliefs^p are better than false ones because true beliefs^p enable us to fulfil our goals better [110].

In the context of ISA$_{bdi}$ our questions can be expressed as follows: What do we mean with the term *epistemic value*? What is the role of *epistemic value* in the environment of ISA$_{bdi}$? Why is knowledge^p more valuable than justified belief^p and why is justified belief^p more valuable than mere belief^p?

There are several different kinds of values, which can be attached to knowledge^p, justified belief^p and belief^p. It is commonly accepted that true belief^p is often instrumentally valuable. Something has instrumental value if and only if it is valuable for the sake of something else meaning it is valuable as a means to some end [67]. But one of the key questions is as follows: *Is true belief^p—as well as justified belief^p and knowledge^p—intrinsically valuable, that is valuable for its own sake in the context of ISA$_{bdi}$?* One way to try to answer this question is to consider whether ISA$_{bdi}$ has an intellectual interest in a truth, which is grounded in ISA$_{bdi}$'s 'curiosity'. Thus, true belief^p would be valuable for its own sake when it answers such an interest [23]. Has ISA$_{bdi}$ such a kind of curiosity? We argue that this is not the case. Even though for some peculiar reasons ISA$_{bdi}$ could be designed and implemented to act as 'curiously' for its own sake, ISA$_{bdi}$ would not have a real kind of motivation to act 'curiously' as ISA$_{bdi}$ is not a naturally curious being.[10] In addition, we claim that in the context

[10]This is the case at the time of writing this thesis.

of ISA_{bdi} all the epistemic entities have value only as a means to provide ISA_{bdi}'s customers with the best possible services (ISA_{bdi}'s existence is motivated only through its capabilities to serve others); thus, ISA_{bdi} is not interested in knowledge for its own sake, but only as the instrumental value.

There could also be eudaemonic value [67]. Eudaemonic value for S is value vis–a–vis well–being of S, in other words, what is good or bad for S [67]. As eudaemonic value is seen as a subjective thing—from the viewpoint of an epistemic agent—we claim that in the context of ISA_{bdi} there is no such value because there is no such psychological phenomenon of ISA_{bdi}, which can be assumed to exhibit well–being of ISA_{bdi}.[11] Therefore, we see the value of knowledge from the perspective of a sui generis domain of epistemic value, distinct from the domain of eudaemonic value.

We continue the discussion about epistemic value with the help of the medical scenario introduced in Section 2.1.1, which we analyse in order to point out our understanding of the epistemic value.

Therefore, we need to evaluate the epistemic value of the information stated by each proposition:

A: *The correct diagnosis is trimalleolus fracture.*

B: *The correct diagnosis is bimalleolus fracture.*

C: *The correct diagnosis is lateral malleolus fracture.*

and how they affect the process of Matti's medical care.

This scenario indicates that the instrumental value of knowledge^p is higher compared to the instrumental value of either justified belief^p or mere belief^p for the proper medical care and best possible recovery. This is because belief^p and justified belief^p are more likely false than knowledge^p, and a right diagnosis is the most essential factor in the determination of correct medical procedures. Thus, we can reason that because ISA_{bdi} is required to achieve its goals, then it would be preferable for ISA_{bdi} to have relevant knowledge^p. The scenario also indicates that reliabilism is a proper approach to knowledge^p and justified belief^p, as the most reliable process (operation) to do the diagnosis leads to the best practice and result. This is because knowledge^p comprise the existence of a reliable^p connection to truth.

In our scenario above the epistemic value of knowledge^p is discussed comparing the epistemic value of the information stated by proposition A

[11] There is a question about what we mean with the term *well–being of* ISA_{bdi}. If its explication comprises the existence of ISA_{bdi}, then there can be eudaemonic value.

(knowledge^p) to the epistemic value of the information stated by proposition B (justified belief^p) and the epistemic value of the information stated by proposition B to the epistemic value of the information stated by proposition C (mere belief^p). The much debated epistemic value problem—the Meno problem by Plato and its derivatives—discusses possible added value of knowledge^p compared to true belief^p in the case when a proposition is the same one. The first derivative is as follows [110]: why is knowledge^p more valuable than any proper subset of its parts? The second derivative requires to explain what special kind of value is achieved once belief^p is transformed from true justified belief^p to knowledge^p? Thus the question is: what is the relevance of these problems in the context of ISA$_{bdi}$?

Duncan Pritchard expresses the Meno problem (the swamping argument) as follows [109]:

P1 *If the value of a property possessed by an item is only instrumental value relative to a further good and that good is already present in that item, then this property can confer no additional value to that item.*

P2 *The value of the property of being a reliably formed belief is instrumental value relative to the good of true belief.*

C1 *Reliably formed true belief is no more valuable than mere true belief. From (P1) and (P2)*

P3 *Knowledge is reliably formed true belief.*

C2 *Knowledge is no more valuable than mere true belief. From (C1) and (P3)*

At first, we are of the opinion that in the context of ISA$_{bdi}$ knowledge^p is not always more valuable than true belief^p, but there are many cases where instrumentally knowledge^p is more valuable than mere true belief^p. Mere true belief^p is more likely to be lost than knowledge^p, which is more stable. Knowledge^p is not entirely stable either, but justified belief^p and mere belief^p are more unstable than knowledge^p. There is a good reason why knowledge^p is more stable than mere true belief^p because knowledge^p, unlike mere true belief^p, could not easily be mistaken [110].

Linda Zagzebski states that the reliability^p of the source of a belief^p cannot explain the difference in value between knowledge^p and true belief^p, if truth is all that matters because reliability^p in itself has no value or disvalue [169]. However, according to her, knowledge^p should not be understood in itself as a state consisting of a known belief^p, but rather as a state which

consists of both the true belief^p and the information of the source of the true belief^p. As Alvin Goldman[12] and Erik Olsson state in their article *Reliabilism and the Value of Knowledge* [58]: *"A reliable^p process itself has value, which can be added to that of the resulting true belief^p to yield a compound state of affairs*[13] *(a knowledge^p state) with more value than the true belief^p alone."* When a true belief^p is produced by a reliable process, the compound state of affairs has a certain property that would be missing, if the same true belief^p were not so produced [58]. Alvin Goldman and Erik Olsson argue that the property of making it likely that one's future beliefs^p of a similar kind will also be true is such a property (conditional probability solution) [58]. Stability is the key component. The extent to which a knowledge^p state enhances the conditional probability of future true beliefs^p depends on a number of empirical regularities [58]. This is a valid argument in the context of ISA$_{bdi}$, as usually ISA$_{bdi}$ processes same algorithms over and over again. There is a connection to the dependability theory of computer science. On the other hand, Duncan Pritchard argues that this reliability^p is a value that attaches to a process producing reliably^p true belief^p and not to true belief^p itself. However, we do not agree that this statement invalidates the above argument in the context of ISA$_{bdi}$ because in the presence of fallibilism and vagueness of the concept of truth the instrumental value of reliably^p produced true belief (knowledge^p) is higher than unreliably^p produced true belief^p. This is linked to the trustworthiness of a true belief^p when ISA$_{bdi}$ uses it in decision making processes, as a reliable^p process attaches its trustworthiness to the true belief^p.

Thus, the question is the following: Do benefits achieved by knowledge^p overwhelm the efforts that are required to achieve knowledge^p?[14] In addition, we argue that in pragmatic circumstances ISA$_{bdi}$ is not interested in whether knowledge^p A is more valuable than true belief^p A, but ISA$_{bdi}$ is actually interested in whether knowledge^p A is more valuable than justified belief^p B and whether justified belief^p B is more valuable than mere belief^p C. This is the case because the answer to the latter questions are more important as factors when deciding on actions to be carried out—just like our scenario above indicates. ISA$_{bdi}$ is not interested in knowledge^p for its own sake, but to provide its customers with dependable services.

Finally, we agree with John Hyman who expresses the difficulty of the knowledge value problem in the following way [71]: *"Instead of asking what we need to add to belief to get knowledge, or how knowledge differs from*

[12] Alvin Goldman himself does not support this.

[13] A compound state consists of a reliable process followed by a true belief^p.

[14] This is an application dependent factor; thus, the problem itself is outside the scope of this thesis.

belief, we are forced to ask how knowledge gets exercised or expressed—since this is invariably how abilities are defined. ..... Knowledge is the ability to be guided by the facts."

3.3 Truth

One of the first definitions of truth is Aristotle's *"For to state of that which is the case that it is not the case or of that which is not the case that it is the case is false, and to state of that which is the case that it is the case and of that which is not the case that it is not the case is true."* [69]. Since Aristotle's time several truth theories have been developed, such as the coherence theory (Francis Bradley) [167], the pragmatic theory (Charles Peirce, John Dewey, Michael Dummett, William James) [53], the correspondence theory (Bertrand Russell, Ludwig Wittgenstein, John Austin) [88, 121], the semantic theories (Alfred Tarski, Donald Davidson) [53], the redundancy theories (Frank Ramsey, John Mackie, Nuel Belnap, Peter Strawson) [143].[15] Because of the many theories of truth there has been an opinion that not all declarative sentences in all domains are true in exactly the same way. This is called *pluralism about truth*. The basic interpretation includes statements that there is more than one truth property, some of which are had by all true sentences [104].

Truth theories can also be classified as 1) *deflationary theories*, which include the redundancy theory, the prosententialism theory, the disappearance theory, the disquotational theory, and the minimalist theory and 2) *inflationary theories*, which include the correspondence theory, the coherence theory, and the pragmatism theory. The key difference between deflationary theories and inflationary theories is the question whether or not truth is a substantive property. Deflationists reject the idea that truth is the substantive property while inflationists support the idea of the substantive property. In other words the disagreement is over the following claim: *there exists some property F (e.g. correspondence) such that any proposition, if true, is so in virtue of being F and this is a fact that is not transparent in concept of truth.* So the inflationary theory of truth claims that *F is necessary and sufficient for explaining the truth of any true proposition p* [104]. According to the deflationary theory of truth, to assert that a statement is true is just to assert the statement itself [143].

As there is not after 2000 years of studies and discussions an unambiguous definition of truth, and as Jonathan Ichikawa et. al. [72] expresses

[15] In the parenthesis philosophers who supported the theory.

"Something's truth does not require that anyone can know or prove that it is true. Not all truths are established truths. ... Truth is a metaphysical, as opposed to epistemological, notion: truth is a matter of how things are, not how they can be shown to be. So when we say that only true things can be known, we're not (yet) saying anything about how anyone can access the truth." clearly point out the difficulty of the issue.

As most epistemologists are of the opinion that what is false cannot be knowledge^p, the concept of truth may play a role when ISA$_{bdi}$ processes information on behalf of human beings in the environment of DIDS. Below we discuss some main features of truth theories related to this thesis. We start with some requirements of truth theories.

Hannes Leitgeb has outlined in his article *What Theories of Truth Should be Like (but Cannot be)* [86] several requirements for a good theory of truth. Below we go through some requirements, which are relevant in the context of ISA$_{bdi}$.

1. **Truth is to be expressed by a predicate and a theory of syntax should be available.**
"There is almost unanimous agreement that truth is to be expressed by a predicate of the form 'is true'— briefly,'Tr'—and thus by a linguistic device that is applied to singular terms which are meant to denote the very objects that are true or untrue. For example, if 'Tr' is a predicate of declarative sentences, then we want to concatenate it with proper names, definite descriptions or variables that refer to these sentences. [86]"

In the context of ISA$_{bdi}$, where truth is usually expressed using artificial languages, there are several different ways to express truth: as a parameter, as a predicate, and as metadata. Which one is best is most often evaluated according to other factors than the concept of truth itself. The factors are more related to expressiveness, purpose, parsing, etc. of the language.

We argue that in the context of artificial languages propositions are more appropriate truth bearers than sentences. In the world of ISA$_{bdi}$ propositions are generally understood as making meaningful claims about what the world is like. Sentences are more connected to the theory of artificial languages: what is a correct syntax, how to parse a sentences, etc.

The predicate *'is true'* or *'Tr'* can be expressed in metadata describing the world where the proposition is stated. The next example shows the use of RDF–language (XML–representation) to express the proposition *"Snow is white at Ivalo"* to be true:

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:ma="http://www.example.org/Materials/">
<rdf:Description rdf:about="http://www.example.org/ColorOfMaterial/" >
    <ma:material>snow</ma:material>
    <ma:color>white</ma:color>
    <ma:location>Ivalo</ma:location>
    <ma:truthpredicate>isTrue</ma:truthpredicate>
</rdf:Description>
</rdf:RDF>
```

2. **If a theory of truth is added to mathematical or empirical theories, it should be possible to prove the latter true.**
"A theory of truth should be designed in a way such that if truth is to be explained for the language of a certain theory T, then adding such a theory of truth to T should allow us to prove (the members of) T true, or otherwise this theory of truth would be either useless or flawed. [86]"

In the context of ISA_{bdi} dealing with logical and mathematical theories this is a rather uncontroversial issue. However, when ISA_{bdi} deals with empirical theories this is a very strong claim. And most often in the context of ISA_{bdi} truth is applied to an 'ordinary' belief; thus, we can deploy here Frapolli's epistemic objections [43]. The theory of truth has to explain the connection between truth and the criteria for its application. When truth is attached to a propositional content, the content is put forward to be used for further assertive acts, inference, or an action affecting ISA_{bdi}'s world. ISA_{bdi} has to be in a position that enables it to believe that the content deserves its support. In order to be in a position to declare *that p* is true, ISA_{bdi} has to check or prove *that p*. Therefore, the meaning of truth does not include any epistemic trait [122].

3. **The truth predicate should not be subject of any type restrictions.**
"If we agree that the sentence '2 + 2 = 4' is true, it is a minor step to admit that also the sentence 'Tr('2 + 2 = 4')' is true. Accordingly, we want to claim that 'Tr('Tr('2 + 2 = 4')')' is true, and so forth. This leads us to higher and higher levels of reflection, but there is nothing obviously wrong about this fact. [86]"

Alfred Tarski tried to solve semantic paradoxes by suggesting a type-theoretic hierarchy of object languages, metalanguages, metametalanguages,

etc. Each of these language levels would have its own truth predicate that is different from the truth predicates of the other levels. However, according to Hannes Leigeb there are compelling reasons for using a simple untyped truth predicate [86]. And Saul Kripke has pointed out that our language contains just one word "true", and in addition there are applications of truth predicates in everyday language for which we would not even know what types should be assigned to them [78]. We are of the opinion that in the contexts of ISA_{bdi} and DIDS the role of those semantic paradoxes do not overcome the benefits of using using a simple untyped truth predicate. As already mentioned above most often in the context of ISA_{bdi} truth is applied to an 'ordinary' belief^p. Therefore, even though in the use of artificial languages, the truth predicate should not be subject to any type restrictions.

4. **Truth should be compositional.**
"Suppose a sentence is built up from other sentences: whether or not this complex sentence is true seems to be determined solely by whether or not its syntactic constituent sentences are true and by the way the latter are put together. This phenomenon is usually subsumed under the umbrella term 'compositionality'; compositionality principles for truth, reference, meaning, and so forth, are among the fundamental principles of semantics. [86, 122]"

In the context of ISA_{bdi} this requirement does not raise any concerns.

5. **The theory should allow for standard interpretations.**
"Speaking of the truth of a sentence without fixing an interpretation of the linguistic expressions within the sentence does not make much sense; without such an interpretation a sentence is not more than a sequence of meaningless signs arranged in accordance with a set of recursive rules. Usually, when we use a sentence we automatically assign an intended interpretation to it. [86]"

In the context of ISA_{bdi} we see this requirement to be mandatory because ISA_{bdi} in the environment of DIDS must obey its intended ontological commitments. Otherwise ISA_{bdi} may not work according to its specifications or the specifications are not defined properly corresponding with requirements of dependable computing.

There are several interesting questions regarding the concept of truth in the context of ISA_{bdi}'s:

1. Is truth the same kind of concept in the context of ISA_{bdi} as it is in the context of human being?

2. What is the role of truth in the context of ISA_{bdi}?
3. Which truth theory or concept would be most appropriate in the context of ISA_{bdi}?
4. What is the truth bearer in the context of ISA_{bdi}?
5. How is truth communicated between ISA_{bdi}s?
6. As theories are instruments, is there any such theory of truth that has significant contribution to the context of ISA_{bdi}?

3.3.1 Truth Theories

In this section we discuss various truth theories and their possible applicability to the contexts of ISA_{bdi} and DIDS.

The Coherence Theory of Truth

According to the coherence theory of truth a belief^p is true if and only if it is a part of a coherent system of beliefs^p [53]. In the context of ISA_{bdi} one of the key questions is the following one: what does coherence actually mean, especially in the cases where ISA_{bdi} is dealing with common–sense beliefs^p formed from perceptions or via testimony. Is it a logical or orderly consistent relationship of beliefs^p, or is it the lack of semantic contradiction between beliefs^p, or is it maximizing satisfaction of constraints between beliefs^p? There is no single, general answer to these questions. In context of DIDS this may create a problem because each epistemic agent operating in DIDS may implement its own interpretation of coherence. This may lead to a situation where the coherent systems of each epistemic agent are not coherent as a whole. For example, an epistemic agent in DIDS can have a fully coherent system of beliefs^p, but this system is totally incoherent with the other epistemic agents' coherent systems in DIDS. Therefore, this may create a situation, where an epistemic agent believes a belief^p to be true, but other epistemic agents do not believe it to be true based on their own coherent systems of beliefs^p. This is a logical incompatibility between the epistemic agents operating in DIDS, and it breaks the law of contradiction. Therefore, the coherence theory of truth creates design, implementation, and dependability problems.

We claim that the coherence theory of truth falls short in the contexts of ISA_{bdi} and DIDS due to the ambiguous nature of coherence, the possibility to break the law of contradiction, and the problems in communications of

true beliefs^p (propositions) between agents in DIDS. And it is possible to establish a coherent world without any correspondence to the real world as we perceive it.[16]

The Pragmatic Theory of Truth

According to pragmatism the idea of truth belongs only to the domain of science, as C.S. Peirce expressed it [84, 103]: *"The ideas of truth and falsehood, in their full development, appertain exclusively to the scientific—experiential—method of settling opinion."* Peirce defined truth to be *"the opinion which is fated to be ultimately agreed by all who investigate, is what we mean by the truth, and the object represented in this opinion is the real."* Pragmatism provides an account of the relations between the concepts of truth, belief^p, and inquiry. According to William James [76, 84] *"The true is the name of whatever proves itself to be good in the way of belief^p, and good, too, for definite assignable reasons. 'The true', to put it very briefly, is only the expedient in the way of our thinking, just as 'the right' is only the expedient in the way of our behaving."* One interpretation of the pragmatic truth theory is that beliefs^p are made true by the fact that they enable us to make accurate predictions of the future run of experience.

We claim that the definitions of pragmatic theory of truth are not exact enough. For example, William James statement above allows every ISA$_{bdi}$ to have its own interpretation and implementation of truth depending on its environment and objectives. This, in turn, may create an environment of DIDS, where one ISA$_{bdi}$'s truth is not another ISA$_{bdi}$'s truth despite same perceptions and beliefs^p. This may affect negatively the dependability of IDS. Hence, the pragmatic theory of truth is either too difficult or leave room for too many different interpretations to be modelled and implemented properly for the environments of ISA$_{bdi}$ and DIDS. In addition, the concept of truth is widely used in everyday communications outside the domain of science. Therefore, we claim that the pragmatic theory of truth falls short of being utilized in the contexts of ISA$_{bdi}$ and DIDS.

The Redundancy Theory of Truth

The basic idea of the redundancy theory of truth is that asserting *that p is true* is completely equivalent to asserting *that p* itself. According to

[16] Though we see that there may exist virtual worlds, such as games, implemented using ISA$_{bdi}$s, where the coherence theory of truth is quite feasible. In these cases a virtual world can be fully coherent and yet a single belief^p is not true outside the world of the game.

redundancy theorists truth is a redundant concept, and the word *truth* does not point to anything in reality [143]. In the contexts of ISA_{bdi} and DIDS the problem with the redundancy theory of truth is that it deals mainly with philosophy of language; therefore, it does not provide a proper theory to implement a solution to infer the truth–value of a proposition. Therefore, we claim that the redundancy theory of truth falls short of being utilized in the contexts of ISA_{bdi} and DIDS.

The Correspondence Theory of Truth

The fundamental idea of the correspondence theory can be expressed as the following ontological thesis: *a belief^p is true if there exists an appropriate entity—a fact, a situation and a type of situation—to which it corresponds* [31, 53]. Thus, we see that truth is a world–to–world[17] relation in the context of ISA_{bdi}, that is ISA_{bdi}'s world models (social model, world model, and mental model) correspond to the world, where ISA_{bdi} operates. We also claim that from the viewpoint of ISA_{bdi} realism is closely related to the idea of truth. In the context of ISA_{bdi} the key features of realism can be expressed in the following way [53]:

1. The world exists objectively, independently of the ways that ISA_{bdi} models or describes the world[18] and

2. ISA_{bdi}'s beliefs^p are about the world.

Now, the question is what kind of features are required by the world–to–world relation in order it to be truth? We claim that reliabilism explains adequately enough the required features in terms of the truth–conduciveness of ISA_{bdi}'s belief^p–forming process (see Sections 3.5.5 and 3.6.5).

We claim that the correspondence theory of truth is appropriate for the contexts of ISA_{bdi} and DIDS. It can be defined exactly and ubiquitously enough for the specifications and implementations of ISA_{bdi}s and DIDS. It does not raise any additional performance issues because the correspondes between world–to–world is already a factor in the contexts of ISA_{bdi}s and DIDS.

[17] In the case of human beings truth is a mind–to–world relation.

[18] The worlds created by virtual reality and augmented reality applications pose severe problems to realism and we are of the opinion that those problems are worth of thorough studies.

The Identity Theory of Truth

The basic idea of the identity theory of truth states that *the content of a declarative sentence is true just if it is (identical with) a fact* [14, 51]. And a fact is defined to be, very generally, *a way things are or a way the world is.* A variation of this definition is that declarative sentences are called *propositions*, and *all true propositions are identical with facts at the level of reference.* The identity theory of truth tries to make a connection between language and reality.

The identity theory of truth is seen as a response to problems of the correspondence theory of truth, but we see that this theory just shifts the problem to the explication of the term *fact.* What actually is *fact*, and how do we verify that something is a *fact*, for example, in the environments of virtual reality and augmented reality. Therefore, we are of the opinion that the identity theory of truth does not bring any benefits (clarity of definition, ease of implementation, etc.) compared to the correspondence theory of truth; therefore, the correspondes theory of truth is more applicable in the contexts of ISA_{bdi} and DIDS.

3.3.2 Speech Act Theory and ISA Asserting Propositions

In order to evaluate the concept of truth in the context of ISA_{bdi}, we need to explore cases where truth plays a role. There are two main components of ISA_{bdi}, where the concept of truth may play an important role. The first one is the inference systems of ISA_{bdi}, where truth is closely connected to various logics, such as modal logics (e.g. the logics of belief, knowledge, time, and action) (see Section 2.2.3). ISA_{bdi} may also operate on natural languages using various artificial languages and logical systems. The second one is the communication system of ISA_{bdi}, where truth deals with both ISA_{bdi}'s assertions to its co–operation partners and ISA_{bdi}'s perceptions from its co–operation partners. As truth is a complex concept, we concentrate only on a sub–area of the concept that is when someone wants to carry out an illocutionary act of asserting a true belief^p. Truth is related to acts of saying something [43]. In the contexts of ISA_{bdi} and DIDS we can see truth–communicative acts as a method of dependable communication. We can compare ISA_{bdi} performing an illocutionary act with human beings performing an illocutionary act, even though ISA_{bdi} does not usually use any natural language to state the proposition but an artificial language developed for ISA_{bdi}–to–ISA_{bdi} information exchange. The illocutionary act of asserting a true belief^p serves to mark out contents that can be

used as premises for inferences or as support for further acts based on the assertion.

To understand the meaning of truth we need to comprehend the characteristics of an act in which truth is ascribed. ISA_{bdi}'s assertions can be modelled using a theory of human–to–human communication called speech act theory, which was developed by J.L. Austin and John R. Searle [127]. We follow the key idea of the FIPA (Foundation for Intelligent Physical Agents) standardization organization to model ISA_{bdi}–to–ISA_{bdi} information exchange [39]. The type of illocutionary act we are interested in is the type including assert, state that, and affirm any proposition p. Standard speech acts in which truth typically plays a role are those that ascribe truth to a propositional content by the use of a specific kind of sentence, a truth ascription, and the truth ascription recovers a proposition already asserted or assumed to be assertable. Truth is a property of things like claims, assertions, beliefs, statements, or propositions; thus, it is not a property of sentences [43]. When we model ISA_{bdi}'s assertions using the speech act theory, we assume that propositions are the truth bearers. Searle defines the assert type of the illocutionary act as follows [127]:

1. Propositional content: *Any proposition p.*
2. Preparatory: *First, S has evidence (reasons, etc.) for the truth of p and second, it is not obvious to both S and H that H knows p.*[19]
3. Sincerity: *S believes p.*
4. Essential: *Counts as an undertaking to the effect that p represents an actual state of affairs.*

Based on this definition ISA_{bdi}'s assertion can be specified as follows [38]:

1. Communicative act: *INFORM*
2. Message content: *Proposition*
3. Description: ***INFORM*** *indicates that the sending agent:*
   a) holds that the proposition is true,
   b) intends that the receiving agent also comes to believe that the proposition is true, and
   c) does not already believe that the receiver has any knowledge of the truth of the proposition.
   From the receiver's viewpoint, the receiver believes that the sender believes the proposition to be true.

[19] S = Speaker, H = Hearer

4. Formal model: < i, inform(j, ϕ) >
   FP: $B_i\ \phi \wedge \neg\ B_i$ ($Bif_j\ \phi \vee Uif_j\ \phi$)
   RE: $B_j\ \phi$
   Where:
   i = agent i; j = agent j;
   ϕ = proposition;
   FP = feasibility precondition; RE = rational effect;
   B = believe; Bif = believe if; and Uif = uncertain if.

An example: ISA_{bdi} *i* informs (asserts) to ISA_{bdi} *j* that (it is true that) snow is white.

(inform
sender: (agent–identifier: name i)
receiver: (agent–identifier: name j)
content: "white (snow)"
language: Prolog)

3.3.3 Thoughts about Truth and ISA_{bdi}

Let us suppose the following hypothetical information exchange between two ISA_{bdi}s S and R:

1. S asserts *"snow is white"* to R.
2. R asks *"how do you know that it is true"* from S.
3. S asserts *"I have knowledge of it"* to R.
4. R asks *"what justifies your belief about it"* from S.
5. S asserts *"I have a reliable perception of it and I obey reliabilism as the knowledge theory"* to R.

Now, what is the role of truth, here? Should we say that in this context truth is **sui generis** or think in the same way as Frege expressed it: *"Truth is obviously something so primitive and simple that it is not possible to reduce it to anything still simpler."* [44] via [43]. Thus, there are no extra benefits to be achieved, at all, by having a more *advanced* concept of truth when we are dealing with ISA_{bdi}'s assertions? So the reliabilism theory of knowledge, as formulated, for example by Alvin Goldman, does the trick by relying on a reliable^p process of knowledge formation (truth–conduciveness of a belief–forming process [57]). Or is truth just hiding in the background as a higher order concept that is not needed explicitly to be taken care of

in the acts of ISA_{bdi}? Or should we agree with Mark Richard's claim about minimalism/deflationism that *the most interesting thing about truth is that it is not very interesting* [118] via [43]. Should we be an inflationist or a deflationist or a some kind of a bizarre combination of both of them when we are dealing with ISA_{bdi}?

We have an intuition that connects truth with a very high degree of epistemic warrant. A truth ascription is an act in which an epistemic agent attributes truth to a propositional content salient in the context [43]. Is there any difference whether the epistemic agent is human being or ISA_{bdi}? We claim that there should be no difference, when for example, ISA_{bdi} acts on behalf of human being.

To understand the meaning of truth it is necessary to understand the characteristics of the act in which a truth is ascribed [43]. We have a very good ground to assume that descriptivism is one of the characteristics. Descriptivism is the view that defends that all declarative sentences describe a state–of–affairs, actual or possible and are thus true or false [43]. As long as human beings are designing ISA_{bdi}, this should be one of the design principles to develop ISA_{bdi} (*the simpler the better* principle). Descriptivism and realism are close to each other. Realism supports the idea that theoretical claims, as describing a mind independent (or world–model independent) world, constitute knowledge^p of the world. The correspondence theory of truth is often associated with the external realism[20] that involves the correspondes principle, according to which truth involves a correspondence between beliefs and external things [25]. The principle of correspondence claims that languages and theories speak and theorize about mind independent entities. According to the correspondence theory of truth, true theories do not aim to copy the world, but aim only at some kind of structural similarity [25]. In the context of ISA_{bdi} this means that the theory according to which ISA_{bdi} builds up its world model, mental model, and social model must enable proper structural similarity between the models and the world in which ISA_{bdi} operates. If this requirement is too strong (too difficult to implement or an implementation would not fulfil performance requirements) then a truth–maker theory could be an alternative approach because it abandons the requirement of the structural relationship, but requires only that true propositions have some worldly truth–maker [25].[21]

Another characteristic of an act is its context–sensitiveness. As Gottlob Frege has advised in his book The Foundation of Arithmetic (page xxii)

[20] As such the correspondes theory does not require its supporters to be realists.
[21] This issue is worth of further study.

"never to ask for the meaning of a word in isolation, but only in the context of a proposition" [45]. We can say that the meaning of ISA_{bdi}'s proposition is always depending on the context where the proposition is asserted. As in contemporary pragmatics, the bearers of meaning and content are not even identified with complete sentences but rather with complete speech acts [43]. This implies that ISA_{bdi}, in addition to asserting a proposition, needs to inform a receiver/receivers of the context where speech acts take place.[22] This can be carried out in the metadata of propositions, which in a way is equivalent to uttering a truth ascription *such–and–such is true 'in this context'* in natural language.

Truth is as much a semantic notion as it is a syntactic and pragmatic instrument [43]. We see that in the context of ISA_{bdi} truth is mostly the semantic notion—unless ISA_{bdi} deals only with logical and mathematical applications—assigning some kind of value of meaningfulness to information that ISA_{bdi} is dealing with. Semantic expressiveness [43] is a view on the type of contribution that a certain term makes to a proposition. The meaning of truth is expressive in the semantic sense. Truth conditions determine what is said. There are the conditions under which a particular content can be asserted, and also the conditions under which truth can be correctly ascribed to it. In the context of ISA_{bdi} this means that ISA_{bdi} must have relevant conditions at its capabilities. That is, in ISA_{bdi}'s world model, social model, and mental model there are required data about the conditions, for example, contextual information, obeyed theories, used algorithms, and truth value.

A very interesting question is as follows: *What does ISA_{bdi} do with truth ascription?* Asserting a content is offering it to others as true, putting forward the content as something that can be used as a premise. Ascribing truth to a content is presenting its status of "usable" in an explicit manner [43]. This thought presented by Frapolli is totally applicable in the context of ISA_{bdi}. When being true a justified belief^p expressed by a proposition is considered to be knowledge^p[23], and knowledge^p provides ISA_{bdi} with a better chance to carry out its responsibility compared to—maybe poorly—evaluated trustworthiness of information providing ISA_{bdi}. Thus, when asserting a truth ascription ISA_{bdi} S provides ISA_{bdi} R with a higher possibility to succeed and to be dependable in its operations.

[22] Note that the receiver is not always aware of the context of the sender; for example, in social media the sender's context is not always known or the context may change from time to time.

[23] There are also other requirements, as discussed in Section 3.6.

3.3.4 Conclusions about Truth in the Context of ISA$_{bdi}$

As truth is in general applied to ordinary beliefs^p—except applications in the domains of mathematics and logic—it is a difficult issue in the context of ISA$_{bdi}$. The question is, does truth as an explicit concept introduce as such any real benefits that motivate to solve the problems? The answer is an application–dependent issue; hence, it is outside of the scope of this thesis. If truth really needs explicitly to be taken into account, we see two possible approaches. First, a minimalist approach to truth "*Tr(snow is white) if and only if snow is white*" and this is all there is to say about the concept of truth could be the appropriate solution for applications that have minimal connections to the real world. Second, the correspondence theory of truth could be the appropriate solution for applications that depend on perceptions from the real world. We argue in Section 3.6 that a form of the reliabilism theory of knowledge^p—as well as justification—is the most appropriate one in the contexts of ISA$_{bdi}$ and DIDS. We are of the opinion that reliabilism scopes truth adequately enough—reliable^p correspondence as the world–to–world connection—in terms of the truth–conduciveness of ISA$_{bdi}$'s belief^p–forming process. The correspondence theory of truth satisfies most of the requirements for truth theories discussed in the article *What Theories of Truth Should be Like (but Cannot be)* [86].

3.4 Belief

In this section we analyse what kind of entity belief^p is in the context of ISA$_{bdi}$. There are two main questions: First, what is the structure of belief^p? And second, what is the role of belief^p?

Belief^p is considered to be an attitude that a human being has whenever she/he takes something to be the case or regards to be true. In addition, in a standard philosophical usage the term belief^p does not have uncertainty about the matter in question. In general, belief^p is characterized to be a propositional attitude, which is the mental state of having an attitude, stance, or opinion about a proposition to be true. Thus, belief^p is the state of having a representation of a proposition stored and believing the proposition to be true [126].[24] At a general level this indicates that ISA$_{bdi}$ has a belief^p when it is in a state, where first, there is a proposition stored (using appropriate representation) in its memory, and second, the truth–value of the proposition is also stored in its memory.[25]

[24]The word *belief^p* is ambiguous between the state and the content of state, i.e. the proposition.

[25]Another option instead of storing the truth value is an algorithm that enables storing

Belief^p is generally considered to play a causal role in the production of behaviour. This is called as a representational approach to belief^p. As one flavour of representationalism Jerry Fodor sees mental representations to be sentences in an internal language of thought (LOT, LOT2): mental representations are structured and they have a compositional semantic [41]. The representational structure is linguistic. A subject believes *that p* only in the case of having a representation of p that plays the right causal role in hers/his/its cognitions. In the context of ISA$_{bdi}$ we see that the language of thought is a prominent approach because various artificial languages, such as Prolog, RDF, and OWL, are already used in the representations of propositions, and representations are structured, stored, and deployed in the processes executed by ISA$_{bdi}$.

There are several other approaches in addition to representationalism such as dispositionalism, interpretationism, functionalism, and instrumentalism. Next we discuss briefly each of these approaches and their applicability in the context of ISA$_{bdi}$.

According to dispositionalism [126] the pattern of actual and potential behaviour is the fundamental thing in belief^p. People having this view of belief^p argue that for someone to believe a proposition *that p* is for that person to possess one or more particular behavioural dispositions pertaining to p. In the context of ISA$_{bdi}$ this raises some design issues. One issue is related to the status of a proposition: is it the object of a belief^p or not? In order a proposition to be the object of a belief^p dispositionalism requires that there is an explicit connection between the proposition and both options (potential behaviour) and actions (actual behaviour) (see Figure 2.6). This creates at least three problems. First, the connection structure may unnecessarily complicate the architecture and implementation of ISA$_{bdi}$. Second, either there are two kinds of propositions in the belief^c database (the models of the world): the objects of beliefs^p and ones that are not associated to any belief^p, or in the belief^c generation phase all the propositions need to be connected to options in order to be stored in the belief^c database. Both these cases are unsatisfactory because the first one requires some kind of an update mechanism and the second one may lead to a situation where important beliefs^c is not stored, at all. Third, an interesting question concerns a belief^c that is stored in the belief^c database of ISA$_{bdi}$, and which has not yet connected to any behavioural disposition, but may be connected in a later phase when ISA$_{bdi}$ learns more about its environment and changes its behaviour. What is the motivation of storing the belief^c (proposition) if ISA$_{bdi}$ does not have any idea of the epistemic

only propositions that ISA$_{bdi}$ considers to be true.

status of the belief^c? How do we track and update possible state changes when ISA$_{bdi}$ learns more? The required mechanisms might be too expensive from the viewpoint of performance requirements. We are of the opinion that these kinds of problems cause dispositionalism to be unsuitable in the context of ISA$_{bdi}$.

Interpretationism [126] is similar to dispositionalism in the way that patterns of actions and reactions are important factors instead of internal representational structures. But interpretationism focuses on observable behaviour, which is interpreted by an outside observer. This approach is even more unsuitable in the context of ISA$_{bdi}$, as it requires a feedback mechanism in order to know whether a stored belief^c is a belief^p or not. Using an external feedback mechanism based on the behaviour of an ISA$_{bdi}$ could be an unnecessarily expensive operation, and the feedback mechanism may have a severe affect on the possibility of ISA$_{bdi}$ to fulfil its performance requirements.

Functionalists argue that mental states, belief^p in particular, are created by their actual and potential causal relations to sensory stimulations, behaviour, and other mental states. There are several causal relationships that can be considered as characteristic of belief^p [126]:

1. Reflection on propositions from which p directly follows, if one believes those propositions, typically causes the belief^p *that* p.

2. Directing perceptual attention to the perceptible properties of things, events, or states of affairs, in conditions favourable to accurate perception, causes the belief^p that those things, events, or state of affairs have those properties.

3. Believing that performing action A would lead to event or state of affairs E, together with a desire to achieve E, will generally cause an intention to do A.

4. Believing *that* p, in conditions favouring sincere expression of that belief^p, will generally lead to an assertion of p.

The first case is straightforward in the context of ISA$_{bdi}$ because ISA$_{bdi}$ usually implements various logics, such as propositional, predicate, and modal logics (knowledge, time, event, etc.). However, there are some open issues. For example, omniscience is one of them, that is, does ISA$_{bdi}$ need explicitly to deduce all the logical results, which it believes, or does ISA$_{bdi}$ believe propositions, which are not stored in its belief^c databases but could be deduced, to be true or false? The second case is equally straightforward because typically ISA$_{bdi}$ should be designed and implemented to utilize such

input mechanisms that provide ISA_{bdi} with adequate reliable^p perceptions of matter and information to form beliefs^p. The third case is the key component of the *belief–desire–intention* architecture, thus, it is self–evident. The fourth case is more difficult because the phrase "*sincere expression of that belief*" is an open issue in the context of ISA_{bdi}. What does "*sincere*" actually mean in the context of ISA_{bdi}? Can ISA_{bdi} by itself be malicious?

Functionalism requires a causal relationship either between the belief^p state and its manifestations in behaviour (forward–looking causal relation) or the belief^p state and the causes of the state in question (backward–looking causal relation), and representationalism requires that belief^p is the state of having such a representation stored [126]. In the context of ISA_{bdi} we can combine these two approaches[26] by stating that functionalism expresses an external, process viewpoint to belief^p and representationalism expresses an internal viewpoint to belief^p. But, there are epistemological issues to be resolved, as it is not the same thing to say *1) to believe is to be in a state that fills a particular causal role* and *2) beliefs are states that represent how things are in a world* [126].

According to instrumentalism belief^p attributions are useful for certain purposes, but there are no underlying facts that people really believe and belief attributions are never in the strictest sense true [126]. We do not see that instrumentalism would be a beneficial approach in the context of ISA_{bdi}, as it would make it difficult to analyse the status of a proposition, which is used in two different purposes. In one purpose a belief^c could be a belief^p (the proposition in question is the object of the belief^p) and another purpose the same belief^c is not a belief^p, but something undefined.

Beliefs^p can be classified to be either explicit or implicit: Proposition p is explicitly believed if the representation of p is actually present in the mind in the right sort of way [126]. In the case of ISA_{bdi} this means that the representation of p is stored in ISA_{bdi}'s belief^c database coded with appropriate semantic language. Proposition p is implicitly believed if the mind does not possess the representation of p [126]. In the case of ISA_{bdi} this means that ISA_{bdi} must somehow infer belief^p *that* p from other beliefs^p. Implicit belief^p faces two problems in the context of ISA_{bdi}. The first one is the omniscience problem and the second one deals with the performance requirements of ISA_{bdi}. Can we assume that ISA_{bdi} has an implicit belief^p which inferring time exceeds the available execution time t on one occasion but does not exceed on another occasion? Therefore, implicit beliefs^p pose

[26] Forward–looking causal relation is more difficult to implement, as it requires some kind of feedback mechanism between the environment of ISA_{bdi} and ISA_{bdi} itself. Thus, we prefer backward–looking causal relation.

problems in the context of ISA_{bdi}.

It is possible to have different degrees of confidence in a belief^p [126]. This actually implies that belief^p may not necessarily be true. Now we can ask whether this is connected to the justification of belief^p or is it somehow irrelevant to justification? If the degrees of confidence is connected to justification, then justification may also have degrees, other than just *not justified* and *justified*. If the degrees of confidence is not connected to justification, then in the context of ISA_{bdi} this implies a requirement of two different belief^p evaluation mechanism: confidence evaluation, which is based, for example, on recommendations and justification evaluation, which is based, for example, on reliabilism. This approach may create severe problems, for example, when the mechanisms do not cohere. We are of the opinion that the degrees of confidence of the belief^p is connected to the degrees of justification for the belief^p.

There are several definitions of belief^p, such as the following ones:

1. *Belief^p is a propositional attitude that takes the proposition in question to be true.*
2. *Belief^p is a propositional attitude that aims at truth.*
3. *Belief^p is a propositional attitude that is individuated by its actual and potential causal relations to sensory stimulations, behaviour, and/or other propositional attitudes.*

When we adapt these definitions to the context of ISA_{bdi}, we need to consider the following two issues: first, the problems with the concept of truth (see Section 3.3), and second, the representation of belief^p. In the context of ISA_{bdi} we assume that propositional attitudes are represented in a linguistic form. We summarize our basic thoughts about belief^p in the context of ISA_{bdi} by the following definition:

Definition. Belief^{pc} is a propositional attitude,

1. which is the state of having an opinion about something to be the case;
2. which is created by its actual and potential causal relations to sensory stimulations, behaviour, and/or other propositional attitudes; and
3. the representation of which—structured if necessary—is stored in a linguistic form.

3.5 Justified Belief

In this section we discuss the concepts of justification and justified belief^p by analysing what is justification and what kind of entity is justified belief^p in the context of ISA$_{bdi}$. We explore various theories of justification, such as foundationalism, coherentism, evidentialism, and reliabilism.

Justification is the key motivation for why an epistemic agent holds a belief^p (proposition) to be true; thus, the role of justification can be seen to help with reaching the truth of a belief^p.[27]

Our argument for ISA$_{bdi}$ to have justified beliefs^p focuses on the scenario presented in Section 3.1 (page 12), which we consider to be representative enough. Using the scenario we argue that if we think a human being having justification for his/her beliefs^p, then we should also think that ISA$_{bdi}$ has justification for its beliefs^p, as well. If we consider that the man in the scenario has the justification for his belief^p *"A pedestrian is crossing the road."* to be true by perceiving via his reliable^p visual capability, then what would be a reason for us to consider that ISA$_{bdi}$ could not have a similar kind of justification for its belief^p *"A pedestrian is crossing the road."* to be true. The perceptual capabilities—their reliability^p—are at equal level in both cases; hence, both recognize at equal level of reliability^p the pedestrian and her/his movement. There is no difference regarding justification in this case.[28] Both epistemic agents have equal capabilities regarding the memory containing the traffic laws; hence, the role of the pedestrian regarding justification is the same. Again, there is no difference regarding justification. The inferring methods of the epistemic agents are not similar, but their reliability^p is at equal level; hence, once again, there is no difference regarding justification. Then, the question is *"Has justification itself a property that is possible only for a human being to manage?"*. Currently we have not recognized any such property. Thus, we argue that ISA$_{bdi}$ can have justifications for its beliefs^p.

When we consider a suitable justification theory for the environment of ISA$_{bdi}$, in addition to traditional epistemological issues we also have a different kind of problems to be resolved, problems of which affect usability and usefulness of a selected justification theory. First, how can we implement ISA$_{bdi}$ that obeys the selected justification theory? Second, what are cognitive requirements of ISA$_{bdi}$ that obeys the selected justification theory? Third, can ISA$_{bdi}$ satisfy its performance requirements[29] when processing

[27] We assume here that belief^p aims at truth.

[28] This is based on reliabilism.

[29] Requirements such as response time, the amount of memory, usability, etc.

justification for beliefs^p and beliefs^p themselves? In addition, there is a question regarding contextualism: is justification dependent upon the context where belief^p is obtained and/or used as a factor in decision–making processes? We are of the opinion that this will be the case because ISA$_{bdi}$s need also to be aware of the consequences of its actions. This will be the case especially in the future where there are ISA$_{bdi}$s that learn new skills.

One of the first problems to which we must find a solution is whether any of the traditional justification theories can be adopted or should a new theory be developed to be utilized in the context of ISA$_{bdi}$. The issue is that in the context of human epistemology a justification theory could be a proper one, but it cannot be implemented using the methods provided by contemporary computer science and AI. Hence, the key factors are related to the capability of implementing a possible justification theory. These factors deal with the issues such as exactness and vagueness of the justification theory and the execution requirements of the implemented theory. The first one means that it may not be possible to make a proper implementation model and specifications based on the theory, or there could be several different implementations of the theory resulting contradictory status of justification.[30] The second one means that even though a selected theory could be implemented its processing requirements (memory, processor, response time, representation languages, logics, etc.) exceed available computing resources.

We start with traditional justification theories. There are several different kinds of definitions of justification (justified belief^p), which could be proper in the context of ISA$_{bdi}$. At first, we discuss internalism and externalism, then we continue with the following topics: 1. *Foundationalism*, 2. *Coherentism*, 3. *Evidentialism*, and 4. *Reliabilism*. The first two ones discuss the structure of justification and the last two ones define the ways of beliefs^p to be justified. We discuss also the role of testimony in exchanging justified beliefs^p between ISA$_{bdi}$s and between ISA$_{bdi}$ and a human being.

There are two notions of justification, which are the doxastic sense of justification and the propositional sense of justification. The doxastic sense refers to the justificational status of belief^p held by a cognizer, and the propositional sense refers to the cognizer's epistemic situation that makes her/him justified in believing a proposition even if she/he does not adopt an attitude of belief^p towards it.

[30] This may raise problems in the environment of DIDS.

3.5.1 Internalism and Externalism

Justification is categorized to be either internalist or externalist. There are several slightly different definitions of internalism. For example, Laurence BonJour defines it as follows [17]: *"A theory of justification is internalist if and only if it requires that all of the factors needed for a belief to be epistemically justified for a given person be cognitively accessible to that person, internal to his cognitive perspective."* Furthermore, Laurence BonJour defines externalism with the help of internalism as follows [17]: *"A theory of justification is externalist, if it allows that at least some of the justifying factors need not be thus accessible, so that they can be external to the believer's cognitive perspective, beyond his ken."*

Internalism and externalism have raised a lot of discussions, where supporters of both approaches have tried to prove their ideas to be the valid ones and the opposers' ideas to be the invalid ones. Critics on internalism point out that most of the problems with internalism arise from the knowability constraints. Strong internalism, which restricts justifiers to conscious states, is stuck with the problem of stored beliefs^p. Weak internalism, which allows stored belief^p as well as conscious beliefs^p to count as justifiers, faces the problem of forgotten evidence and the problem of concurrent retrieval [60]. We see these critics on internalism to be valid also in the context of ISA$_{bdi}$.

In the case of weak internalism Alvin Goldman points out in his article Internalism Exposed [59] a problem, which is severe also in the case of ISA$_{bdi}$. He calls this problem *the problem of concurrent retrieval*. In weak internalism only conscious and stored mental states are justifiers, but it does not express that all sets or conjunctions of such states qualify as justifiers. If a certain set of stored beliefs can all be concurrently retrieved at time t and concurrently introspected, then they could qualify as justifiers under the principle of indirect knowability. But if they cannot all be concurrently retrieved and introspected at t, they would fail to be justifiers. Alvin Goldman claims that concurrent retrieval and introspection is not possible for human beings because such concurrent retrieval is psychologically impossible [60]. Now, in general, the same applies to ISA$_{bdi}$ by requiring an environment, where concurrent retrieval[31] of ISA$_{bdi}$'s epistemic responsibilities might be in strong contradiction with ISA$_{bdi}$'s performance requirements (e.g. response time, processors and memory usages).

Alvin Goldman sees strong internalism as follows [59]: At first, *"The only facts that qualify as justifiers of a person's believing p at time t are*

[31] In the context of ISA$_{bdi}$ concurrent retrieval is not an exact term: at what level of the architecture of ISA$_{bdi}$ system concurrency is thought to be.

facts that the person can readily know by introspection, at t, to obtain or not to obtain." Then, *"Only facts concerning what conscious states the person is in at time t are justifiers of the person's belief at t."* According to Alvin Goldman, this faces the problem of stored beliefs^p. Normally, the majority of the person's beliefs^p are stored in memory rather than occurrent or active. Furthermore, usually in the person's consciousness at the time **t** there is nothing that justifies those stored beliefs^p. Thus, according to strong internalism, then, none of these beliefs^p are justified at the time **t**. This is a major argument against strong internalism. Now, in the case of ISA$_{bdi}$ it could be theoretically possible to store and retrieve at the time **t**[32] all the justifiers of every belief^p of every 'conscious' state of ISA$_{bdi}$[33]; thus, to avoid the problem of stored beliefs^p. However, in practice this would require huge real–time databases of beliefs^c and metadata,[34] which, in turn, would cause severe performance and storage problems because every possible justifier should be verified. And this once again may lead to ISA$_{bdi}$'s epistemic responsibilities being in strong contradiction with its performance requirements.

Thus, we can argue that internalism is not a proper approach to justification in the context of ISA$_{bdi}$.

Externalism [17] does not require that a person whose belief^p is justified has any sort of cognitive access to factors that provides justification. For example, in reliabilism the main requirement for justification is that belief^p must be produced in a way or by a process that makes it objectively likely that belief^p is true. In this case a person has no reason to consider that belief^p is true or likely to be true, but will be epistemically justified in accepting belief^p. In the context of ISA$_{bdi}$ this approach to justification has a significant benefit compared to internalism: The requirements of the cognitive skills of ISA$_{bdi}$ are much lower meaning that an implementation of ISA$_{bdi}$ is far less complicated—the more simple the solution is, the more beautiful it is. Reliabilism is one of the prominent externalist theories. We discuss process reliabilism in more detail in Section 3.5.5, and we argue that it is the appropriate justification theory in the context of ISA$_{bdi}$.

Critics on externalism have judged it to be unsuitable for realizing the true and original goals of epistemology [19]: *"In the end it may be possible to make intuitive sense of externalism only by construing the externalist as simply abandoning the traditional idea of epistemic justification or rationality and along with it anything resembling the traditional conception of*

[32]Time **t** is vaguely defined, here. It is not clear what is the duration of **t**. Is it measured in nanoseconds, milliseconds, or minutes?

[33]For example, future quantum computers may have required performance capabilities.

[34]Of course, this is application dependent factor.

knowledge". We see this critique on externalism not to be meaningful in the context of ISA_{bdi} because our idea of ISA_{bdi} having justified beliefs^p is itself also outside the traditional idea of epistemic justification.

We see that in the context of ISA_{bdi} there are four possible justification theories: foundationalism, coherentism, evidentialism, and reliabilism. Now, we consider each of these theories from the viewpoint of implementing it as the theory of ISA_{bdi}'s beliefs^p justification. In addition, we evaluate testimony in the context of ISA_{bdi}.

3.5.2 Foundationalism about Justified Belief

Foundationalism is a theory of the structure of justification. According to this theory justified beliefs^p form a hierarchical structure, where basic beliefs^p establish a base on which other beliefs^p can be justified. Foundationalism is seen as a solution to the regress problem. A version of foundationalism (Doxastic Basicality) defines [142]: *A person's justified belief that p is basic if and only if the person's belief that p is justified without owing its justification to any of person's other belief.* There are two primary questions: 1. What are methods for ISA_{bdi} to obtain basic beliefs^p? In other words, what is the direct justification of ISA_{bdi}'s beliefs^p? 2. How can ISA_{bdi}'s basic beliefs^p justify ISA_{bdi}'s non–basic beliefs^p? In other words, what is the indirect justification method of ISA_{bdi}'s non–basic beliefs^p? To the first question there are two primary options:

1. Human beings designing and implementing ISA_{bdi}, and
2. ISA_{bdi}'s perceptions.

In the first option the solution would be a "creator's view" to ISA_{bdi}'s justified beliefs^p meaning that human beings designing and implementing ISA_{bdi} provide it with default, basic justified beliefs^p. This, in turn, transfers the requirement of the justification of basic beliefs^p to the next level: How are basic beliefs^p justified to human beings designing and implementing ISA_{bdi}s? This has been discussed in more detail, for example, in the articles [16, 18, 49, 55, 132]. In the BDI architecture discussed in Section 2.2.1 the direct justification means that there should be a default semantic, structured data (propositions and their semantic), and foundationalistic status of belief^p coded into the world model, the mental model, and the social model. This raises a question: If ISA_{bdi} perceives a defeater for a basic belief^p and the basic belief^p looses its justification during runtime, then what should be done? Should the operation be halted, and let designers and implementers resolve the situation and correct the whole belief^p

database or should ISA_{bdi} by itself resolve the situation and update the whole belief^p database during runtime? There is no simple answer. The first one may lead to serious dependability problems, especially reliability^c would be low. The second one may lead to performance problems because the correction of the whole belief^p database could be an expensive, time consuming operation. In the second option we may consider that perceptions are such that they fulfil requirements for basic beliefs^p. Naturally this leads to the question: What are those requirements? For example, we do certainly not consider perceptions from a defective instrument to be basic beliefs^p. One approach could be reliabilism (see Section 3.5.5). Alvin I. Goldman [60] introduced belief–independent processes that could produce justified beliefs^p. These beliefs^p are justified by virtue of being the product of reliable^p processes.

Other proposals for the basic beliefs^p, such as *self–evidence, self–justification, self–warrant, justification by a direct awareness of what a belief is about*, are problematic because those terms are not exact enough for specifying them to be implemented. For example, what kinds of beliefs^p are self–justifying: logical, mathematical, ethical, political, etc. or is it even possible to specify such a categorization? In order to evaluate whether a belief^p is justified by being *self–'something'* actually requires a lot of background understanding of the matter, which may lead to an unnecessarily complicated implementation of ISA_{bdi}, for example, in the cases where learning new skills is required. Therefore, we argue that these options do not seem to be appropriate for ISA_{bdi}.

To the second question one obvious answer could be deductive inference[35](see Section 2.2.3). Therefore, we may consider that these logics may transfer justification from basic justified beliefs^p to non–basic beliefs^p. But the logic of justification is not yet advanced enough, even thought a form of justification logic has been developed [7, 8]. Epistemic logics really work only with belief^p and knowledge^p [8]. Probabilistic inference could also be used, for example Bayesian probability, and classical enumerative induction may also satisfy the requirement [48]. But in the context of ISA_{bdi} the parallel use of several epistemic logics (belief^p, knowledge^p, and justification) might cause the implementation of ISA_{bdi} to be unnecessarily complicated and might also create problems to fulfil ISA_{bdi}'s performance requirements.

Based on the problems presented above we are of the opinion that in general, foundationalism falls short of being the overall theory of the structure of justification in the context of ISA_{bdi}. But there is an exception,

[35] Such as modal logics or other appropriate logics that maintains truth.

namely reliabilism that can be categorized to be a form of foundationalism (see Section 3.5.5). A reliabilist foundationalist can think that basic perceptual beliefs^p are justified by reliable^p sensory experiences or appearances. But there is a question about whether justification could be achieved too easily in the context of ISA$_{bdi}$.

3.5.3 Coherentism about Justified Belief

Coherentism is a theory of the structure of justification. According to this theory there are no basic justified beliefs^p [142]. Justified beliefs^p form a web–like structure, where each belief^p is justified only if it is coherent with other beliefs^p in the system. Hence, every beliefs^p must be coherent with each other in a system formed by this web–like structure. As Keith Lehrer states the role of coherence [85]: *"The input of perception and the output of action supplement the central role of the systematic relations belief^p has to other beliefs^p, but it is the systematic relations that give the specific justification it has."*[36]

There are two forms of coherentism [85]: weak coherence theories and strong coherence theories. Weak coherence theories define that the way, in which belief^p coheres with the background system of beliefs^p, is one determinant of justification; others being perception, memory, and intuition. Strong coherence theories define that justification is solely the matter of how belief^p coheres with the system of beliefs^p.

A version of coherentism (Doxastic Coherentism) defines [142]: *Every justified belief^p receives its justification from other beliefs^p in its epistemic neighbourhood.* This definition raises two important questions: First, what is the epistemic neighbourhood of belief^p in the context of ISA$_{bdi}$? And second, what does coherence mean in the context of ISA$_{bdi}$? The epistemic neighbourhood seems to be an application dependent factor. It can be the structure of ISA$_{bdi}$'s beliefs^p as a whole including the world model, the mental model, and the social model; or it can start from, for example, a sub–part of the world model ending at the structure of beliefs^p of the society of multiple ISA$_{bdi}$s. In addition, there is the question whether the epistemic neighbourhood is static or dynamic throughout ISA$_{bdi}$'s existence. A dynamic epistemic neighbourhood seems to be challenging to implement. The verification of a belief^p being coherent with other beliefs^p in the dynamic epistemic neighbourhood can require so much processing

[36] According to semantic coherentism belief^p has the content that it does because of the way in which it coheres within a system of beliefs^p.

power that it would cause severe problems to ISA_{bdi}'s performance requirements. Therefore, the scheme of the epistemic neighbourhood requires more study, before it is properly understood and could be implemented in the context of ISA_{bdi}.

The second question is even more difficult to cope with. We have already dealt with coherence in Section 3.3.1, where we discussed the coherence theory of truth. Very much same problems are valid in the case of justification. Is coherence a logical or orderly consistent relationship of beliefs^p, or is it the lack of semantic contradiction between beliefs^p, or is it maximizing satisfaction of constraints between beliefs^p? There is no single, ubiquitous solution to these questions. There are some proposals [100], but it is still very much an open issue; unless we interpret coherence purely logically. Coherence can be defined as the quality or the state of cohering—forming a whole—meaning especially a logical or orderly consistent relationship of beliefs^p. But actual beliefs^p quite often do not seem to obey any formal logic. One interpretation is to consider coherence as the lack of contradiction; especially, the lack of semantic contradiction. In the context of ISA_{bdi} we can interpret this as maximizing satisfaction of constraints between beliefs^p. This approach is presented in more detail by Joseph Sindhu [133].

In DIDS the possibility of ambiguous interpretations of coherence may create problems because each epistemic agent operating in DIDS may implement its own interpretation of coherence. This may lead to a situation where a beliefs^p justified by being coherent with other beliefs^p in the belief^p database of one ISA_{bdi} is not coherent (therefore, not justified) with other beliefs^p in the belief^p database of another ISA_{bdi}. This creates a problem at the DIDS level whether the belief^p is justified or not?

Coherence theory may also create an environment, where the epistemic responsibilities of ISA_{bdi} are in conflict with the performance responsibilities of ISA_{bdi}—in particular, this is valid in real–time DIDS—especially, if it requires that every time when a new belief^p is generated coherence has to be verified in order to justify the belief^p. Hence, whenever ISA_{bdi} perceives a belief^p, it needs go through every justified belief^p it has at the moment and verify coherence between the existing justified belief^p and the new belief^p. In addition, what is the role of implicit beliefs^p? Should they be taken into account, as well? If so, it could be an overwhelming operation.

Due to the challenges, that coherentism faces in the context of ISA_{bdi}, we argue that coherentism is not suitable for the general justification theory for ISA_{bdi}.

3.5.4 Evidentialism about Justified Belief

The topic of evidentialism is to define factors based on which an epistemic agent is justified in believing a proposition. Evidentialism is not about when the epistemic agent's believing is justified. The basic idea of evidentialism is expressed in the following definition [96]: *Epistemic agent is justified in believing proposition p at time t if and only if epistemic agent's evidence for p at t supports believing p.*

The main questions of evidentialism are as follows: First, what sorts of things can be considered to be evidence, and second, how can evidence support in believing a proposition? To the latter question our intuition says an epistemic agent must have good, adequate reasons for considering the proposition in question to be true. Hence, the dependence on reasons is central to concept of justified belief [96]. To former question there are many different answers—several different evidentialist theories—depending on how the concept of evidence is explicated. In general, evidence for or against a proposition is any information relevant to the truth or falsity of the proposition. For example, according to the above definition *"only facts that an epistemic agent has are relevant to determining what the epistemic agent is justified in believing meaning that epistemic agent must be aware of, to know about those facts"* [96].

There are objections to evidentialism, such as (1) forgotten evidence, (2) evidentialism, which bases on evidence making a proposition to be probable, is false, and (3) pragmatic reply [96]. The first one deals with the cases in which at first an epistemic agent has had an evidence for a proposition but later on she/he/it has forgotten the evidence, but nevertheless continues to believe justifiably without possessing any other evidence. The second one argues that the possession of reasons that make p probable, all things considered, is not sufficient for p to be justified [96]. The third one is based on William James argument that *having adequate evidence is not necessary for an epistemic agent to believe justifiably. Our hopes, fears, and desires do influence what we believe* [96]. In the context of ISA_{bdi} the first objection is also a valid one. Evidentialism is considered to be an internalist theory, and we have already discussed the case of forgotten evidence related to internalism in Section 3.5.1. The second objection is interesting because if it is true then evidentialism is false also in context of ISA_{bdi}. However, the discussion of truth or falsity of evidentialism is outside the scope of this thesis. The third object raises a question of ISA_{bdi} having hopes, fears, and desires which can influence what ISA_{bdi} believes. We are of the opinion that ISA_{bdi} is not yet capable to have those kinds of attitudes. Hence, the third objection is not valid in the context of ISA_{bdi}.

We argue that evidentialism is not good enough theory to be used in the context of ISA_{bdi}. First, the concept of evidence is not explicated exactly enough and this may lead to various, incompatible kinds justifications in the context of DIDS. Second, evidentialism is considered to be an internalist theory, and we have already argued that externalism is a better approach than internalism.

3.5.5 Reliabilism about Justified Belief

Reliabilism explains important epistemic concepts in terms of the truth–conduciveness of an epistemic agent's reasoning, belief^p–forming processes, methods, faculties, etc. [57]. The epistemic agent's truth–conduciveness is its likelihood to produce true beliefs^p, thus to avoid false beliefs^p. The fundamental idea is that belief^p *that p* is justified on the basis of a reason or ground **r** just in case **r** is a reliable^p indication *that p* to be true [57]. There are reliabilist theories of knowledge^p and justification; as well as reliable–indicator theories and reliable^p–process theories [57]. In this section we concentrate on the reliable^p–process approach to justification. Process reliabilism is able to manage justification of belief^p in both the doxastic and propositional sense of justification [57].

Alvin Goldman argues in his article *What is justified belief?* [60] that reliabilism should specify non–epistemic conditions for the epistemic quality of being justified in order to avoid circularity. Thus, only non–epistemic concepts such as psychological (belief and experience), metaphysical (causation), and relations between propositions (logical deductibility, probabilistic coherence, and degrees of confirmation or support) can be used to evaluate justification for belief^p. A belief^p–generation process that is highly reliable^p—that is, has a high truth–ratio—confers justifiedness on its outputs; hence, a belief^p–generating process that is not highly reliable^p does not confer justifiedness on its output.

Reliability^p can be understood in the frequency sense (pertaining to what occurs in the actual world) or a propensity sense (pertaining both to actual world and possible world outcomes) [57]. What is a reliable^p indication for belief^p to be true in the context of ISA_{bdi}? First, the reason must make the probability of belief^p to be true high in ISA_{bdi}'s normal world[37] [57]. ISA_{bdi}'s normal world is the environment where ISA_{bdi} has been designed and implemented to operate. The probability factor connects

[37] A. Goldman argues that the approach of normal worlds is problematic; thus, this needs to be explored more.

justification to both the dependability theories of computer science and warrant/certification systems in human society. We claim that this connection forms a valid relationship between ISA_{bdi}'s perception (inference, memory) and the justification of formed belief^p (created when processing ISA_{bdi}'s perception, when inferring from existing justified beliefs^p, or recalling from memory). The dependability theories of computer science form the basis to specify the requirement of safety^p. Safety^p can be explained as follows: If ISA_{bdi} believes that p, then p would not easily have been false [107, 139]. Now, using the dependability theories of computer science it can be evaluated whether p would not be easily false. In this case human beings can have "a creator's view" to ISA_{bdi}'s normal world and can specify based on the application requirements the probabilistic limit of safety^p or ISA_{bdi} can itself by learning specify the requirement of safety^p. One approach could be Alvin Plantinga's theory of warrants in the context of human society: *A belief has warrant only if it is produced by cognitive faculties that are functioning properly in an appropriate epistemic environment* [105]. And it is up to a society to decide whether cognitive faculties function properly or not (what is the expected probability to produce correct results) in the appropriate environment.[38]

Alvin Goldman advocates for process reliabilism and he discusses and motivates it in his article *What Is Justified Belief?* [60]. From the viewpoint of ISA_{bdi} he has three important hypothesis:

1. Justification is necessary for knowing, and closely related to it.
2. A theory of justified belief^p shall be specified in non–epistemic terms when a belief^p is justified.
3. There is no such assumption that when a person has a justified belief^p, he knows that it is justified and knows what the justification is, and he can state or explain what his justification is.

All these hypotheses have significant effects on the implementation and operation of ISA_{bdi} including belief^p management, situation management and goal activation, planning, and scheduling activities. One of the critical factors is the requirements of ISA_{bdi}'s cognitive capabilities. The first hypothesis establishes a relation between knowledge^p and justified belief^p, which may help ISA_{bdi} to better evaluate the epistemic quality of information. The second hypothesis enables the evaluation of justification to be based on, for example, probabilities, causal relations, and semantics,

[38] For example, a society has norms to decide when a person has cognitive faculties functioning correctly in the disciplines of medicine, legal affairs, etc.

implementations of which are less complex, because there exist already many examples of their implementations. The third hypothesis—the rejection of knowability requirement—simplifies significantly the structures of world model, mental model, and social model; in addition to processing justification, justified beliefs^p, and beliefs^p. For example, the knowability requirement might cause an ISA$_{bdi}$'s world model to comprise a proposition, a metadata describing the justification status of the proposition, and a metadata of the metadata describing how ISA$_{bdi}$ knows the justification (regress problem). This might create ISA$_{bdi}$'s world, social, and situation models to be complicated. The abandonment of knowability requirement shifts the challenge to the decision making about when there really exists the first level justification for a belief^p. We currently consider that the philosophical regress problem is not so important in the context of ISA$_{bdi}$; hence, we do not require that justification needs justification.

The next important point that Alvin Goldman raises in his article is that if we agree that principles of justified belief^p must make a reference to causes of belief^p, then what kind of belief–forming process is acceptable to justifiedness [60]. We argue that in the context of ISA$_{bdi}$ (as well as human being) reliability^p is the key factor. As Alvin Goldman expresses it [60]: *"The justificational status of a belief^p is a function of the reliability^p of the process or processes that cause it, where reliability^p consists in the tendency of a process to produce beliefs^p that are true rather than false"*. When we are using the term *tendency of process*, we also imply that some beliefs^p can be more justified than others; hence, the degree of justifiedness seems to be a function of reliability^p. In the context of ISA$_{bdi}$ different degrees of justifiedness of various beliefs^p may play an important role, when planning and deciding intentions to be carried out.

Let us have an example of ISA$_{bdi}$ trying to achieve a goal G. There are two different beliefs^p **A** and **B** which lead to different intentions $\mathbf{I}_1$ and $\mathbf{I}_2$ to achieve the goal G:

1. Belief^p **A** (reliability^p: 0.99) $\Rightarrow$ intention $\mathbf{I}_1$ $\Rightarrow$ goal G.
2. Belief^p **B** (reliability^p: 0.55) $\Rightarrow$ intention $\mathbf{I}_2$ $\Rightarrow$ goal G.

According to our intuition and experience more reliably^p created beliefs^p will more likely result in selecting correct intentions^c to be carried out in order to achieve the goal. Therefore, in this example the belief^p *A* should have more weight in the decision making than the belief^p *B*, and the intention $\mathbf{I}_1$ should to be selected to be carried out to achieve the goal G. The higher the degree of justifiedness of a belief^p is, the higher its importance is in decision making.

There is a relation between the degree of justifiedness and the degree of trustworthiness. The question is *how reliable^p must a belief–forming process be in order for beliefs^p supported by it to be justified?*[60].[39] This question is related to contextualism: the criticality of the consequences of ISA$_{bdi}$'s actions affect the reliability^p requirements. This, in turn, requires ISA$_{bdi}$ to be aware of the consequences of its actions.

As a variant of process reliabilism—high probability of truth—is Alvin Plantinga's functionalist theory of warrant. According to this theory, there must be a design plan. A belief^p having warrant requires that the segment of the design plan governing the production of beliefs^p is aimed at truth [56]. We claim that human beings (computer scientist, engineers, system designers, and programmers) can build ISA$_{bdi}$s whose beliefs^p are warranted. This issue can be based on the dependability theories of computer science, which basis is presented, for example, in Jean-Claude Laprie's article *Dependable Computing. Concepts, Limits, Challenges* [83]. We discuss the dependability theories more in Section 2.1.2 and Chapter 5. Of course, the dependability requirements of ISA$_{bdi}$'s application—what are the actual uses of justified beliefs^p and consequences of using unjustified beliefs^p—set the ultimate reliability^p requirements for the belief–forming process or processes to produce justification.

There are three main sources of ISA$_{bdi}$'s perceptions: 1. sensors, 2. other ISA$_{bdi}$s, and 3. human beings. In addition to these three external sources, ISA$_{bdi}$s can obtain justification for its belief^p by an inferring process. Each of these sources form quite a different environment to evaluate the probability of a belief^p to be true.[40] In the case of sensors the reliability can mostly be computed, for example, based on the data provided by equipment manufacturers, dependability theories of computer science, and previous observations of the behaviour of sensors. The second and third case are discussed in more detail in Section 3.5.6.

Based on this brief discussion we argue that in the context of ISA$_{bdi}$ a version of process reliabilism is the most proper justification theory in the group of traditional justification theories. It is an intuitive way of thinking justification of beliefs^p in the context of ISA$_{bdi}$. In addition, process reliabilism is able to manage justification of belief^p in both the doxastic and propositional sense of justification [57]. It does not create similar kinds of design and implementation challenges like other forms of foundationalism and coherentism do; hence, its implementation is the least problematic. And we claim that process reliabilism will not cause such performance

[39] The Sorites paradox causes severe challenges to specify any valid reliability^p value.
[40] Each of these sources need to be evaluated separately, which is a future topic.

problems like the other justification theories do because it more close to the dependability concepts of computer science and their implementations (e.g. correctness of actions) compared to other justification theories.

3.5.6 Testimony about Justified Belief

In multi–agent systems ISA_{bdi}s do not usually operate in isolation, but they establish a (social) network of co–operating ISA_{bdi}s and human beings. Hence, other ISA_{bdi}s and human beings are important sources of ISA_{bdi}'s beliefs^p. Testimony has caused a lot of discussion in epistemology during the last decades. The problem of testimony is related to the problem of justification: What makes a receiver of a belief^p justified in accepting the belief^p that a sender has asserted [1]. There is a kind of default rule for testimony: *If the speaker S asserts that p to the hearer H, then, under normal conditions, it is correct for H to accept (believe) S's assertion, unless H has special reasons to object* [1]. In this rule *accept* can be considered to be a short form of *acceptance as true*. This can be connected to the knowledge norm of assertion that states: *One correctly asserts that p only if one knows (or represents oneself as knowing) that p*[41] [1].

Another question is that what is the real role of testimony: is it the actual source of justified beliefs^p for the receiver or is just a method to exchange justified beliefs^p between the sender and the receiver? If testimony is the source of justified beliefs^p then what justifies testimonially–based beliefs^p? If testimony is the method of exchanging justified beliefs^p, then first, is there a requirement that the sender is allowed to transmit only justified beliefs^p (or knowledge[42]), and second, how is justification transmitted from the sender to the receiver?

There are several attitudes toward testimony, such as reductionism versus anti–reductionism and inferentialism versus non–inferentialism. Reductionism is similar to inferentialism and anti–reductionism is similar to non–inferentialism. The key difference between inferentialism and reductionism is that inferentialists rely only on inferentially–based justification and reductionists allow other types of justification to be taken into account. Reductionists argue that testimony–based justification is based on a combination of inferentially–based, memorially–based, and perceptually–based justification; thus, it is not actually one of its own kind [62]. In other words,

[41] There are many objections to and variations of these principles, such as *Acceptance Principle, Principle of Charity, and Co–operative Principle.*

[42] This requirement has been set, for example, by John R. Searle in his speech act theory and in the communicative act specifications of FIPA.

reductionists believe that the justification of a belief^p is the result of other confirmed sources.

Michael Dummett [35] argues that *"Testimony should not be regarded as a source, and still less as a ground, of knowledge: it is the transmission from one individual to another of knowledge acquired by whatever means."* If testimony is not the source or ground of knowledge^p, then can it be the source or ground of justified belief^p? We consider that it is not possible because we are of the opinion that there is no such factor F, which would make such a distinction between justified belief^p and knowledge^p that testimony could be the source or ground of justified belief^p but not the source or ground of knowledge^p. In the context of ISA$_{bdi}$ reductionism implies that the testimonially received belief^p should be processed in the same way as other perceptions. As we support a form of process reliabilism this indicates that the justification of the testimonially received belief^p is based on the reliability of the belief–forming processes that produced belief^p at an original source of the justified belief^p, the reliability^p of transmission media, and the reliability^p of belief^p perception processes of the receiver (see Section 3.5.5).

Anti–reductionists argue that testimony–based justification is based on the reliability^p of an innate cognitive feature of the mind, which causes us to trust people who testify [62]. In the case of anti–reductionism we have to ask what could be *the innate cognitive feature* of ISA$_{bdi}$, if ISA$_{bdi}$ has any? For example, could it be an algorithm with which the trustworthiness of collaborating partners is evaluated? Anyhow, the term *innate cognitive feature* is very vague; hence, we are of the opinion that it does not form any sound base for justification to be implemented in the context of ISA$_{bdi}$.

Inferentialists consider testimony–based belief^p to obtain justification through the following argument [62]:

1st: S informs R *that* p;
2nd: S has generally been reliable in the past when informing things like p;
3rd: S is probably reliable on this case;

Conclusion: For R p is justified belief.

As a related issue in computer science the evaluation of the trustworthiness of S has gained a lot of interest, especially in the domain of commerce in the Internet, and several solutions have been proposed [120]. There is an open question: how can justification for trustworthiness be obtained in

the case, where there is no past history? In the context of ISA_{bdi} we see inferentialism to be a part of justification as a supporting factor.

Non–inferentialists consider that S informing *that* p itself can be adequate for R to have p as a justified belief^p. Thus, non–inferentialism leads to the requirement that the sender S is allowed to transmit only justified beliefs^p (knowledge^p), otherwise the sender S is malicious. We consider this not to be an ideal approach to the exchange of beliefs^p between ISA_{bdi}s and between human beings and ISA_{bdi}s in the environment of DIDS. Because this is not realistic in the actual world, where there are many malicious sources of information. However, there are models that rely on the premise that ISA_{bdi} is allowed to transmit only knowledge^p, for example, the FIPA agent communication model, which is based on John R. Searle's speech act theory [38].

Robert Audi [9] argues that testimonially–based beliefs^p are formed directly, but they are justified on the basis of other beliefs^p in a way that the other beliefs^p only support the testimonially–based beliefs^p and are not necessarily closely linked to the supported beliefs^p. Jennifer Lackey [79, 80] argues that R comes to know *that* p via S's statement *that* p only if (i) S's statement *that* p is appropriately connected with the fact *that* p; and (ii) R has no defeaters indicating the contrary. Jason Stanley [141] suggests that the amount of required certainty may affect justification; thus, the level of stake should be taken into account when evaluating the requirement of justification.

When we consider the role of testimony in the context of ISA_{bdi}, we need to separate the transmission of justified belief^p into components, whose roles in justification is analysed below. There are five main components involved, when a justified belief^p is transmitted from a sender S to a receiver R:

1. Justified belief^p *JB*,
2. Sender of belief^p *S*,
3. Possible intermediate proxy distributors *PD*,
4. Receiver of belief^p *R*, and
5. Transmission media *TM*.

We have already discussed justified belief^p above, but in this context we see that the important questions related to the first component *JB* are the following ones: What kind of entity is justification in the context of ISA_{bdi}? How is justification expressed? The key question in the context of the second component S deals with justification: What is the status

of the justification for the sender's belief^p or is it actually needed? The role of the third component *PD* raises a question: Does she/he/it need to have justification for the belief^p or can she/he/it be ignorant of justification? The fourth component R is the most interesting one from our point view. And it is also most discussed having debates on reductionism versus anti-reductionism and inferentialism versus non–inferentialism. The fifth component *TM* is not discussed much in epistemology, as it seems to be assumed that the content of a justified belief^p is not altered during transmission, and the semantics of the justified belief^p remains the same in the contexts of both S and R.[43]

There are two dimensions to consider what kind of entity justification is: deontological versus non–deontological and evidence versus reliability. According to deontological justification *S is justified in believing that p if and only if S believes that p while it is not the case that S is obliged to refrain from believing that p* [142]. In the context of ISA$_{bdi}$ this approach is difficult because the meaning of *to be obliged* is quite ambiguous: does it mean that *that p* is not coherent with another justified belief^p, or does it mean that the performance of ISA$_{bdi}$ is not adequate to carry out the verification of *to be obliged*, or does it mean that there is a defeater? This may create a situation where there are several different kinds of justification. Therefore, we argue deontological justification is not suitable in the contexts of ISA$_{bdi}$ and DIDS. According to non–deontological justification *S is justified in believing that p if and only if S believes that p on a basis that properly probabilifies S's belief that p* [142]. As non–deontological justification is closely linked to reliabilism when specifying *properly probabilifies*, we consider it to be appropriate justification in the contexts of ISA$_{bdi}$ and DIDS. *Properly probabilifies* can also be linked to the dependability theory of computer science.

The difference between requiring evidence for justification or reliability^p for justification is that when both of them require some kind of warrant for justification, reliability^p for justification requires that a belief^p is justified if and only if it is a result of a reliable^p, cognitive origin [142]. But what does *cognitive origin* mean in the context of ISA$_{bdi}$? Can we say that ISA$_{bdi}$ transmitting a belief^p stating that *'The temperature is 8 degrees Celsius.'* exhibits a reliable cognitive origin? We consider that reliability^p can be derived from the dependability theory of computer science, and the reliability^p of the thermometer (based on the certificate of the manufacturer), and *cognitive origin* can be derived from artificial intelligence. We

[43]Note! Hearing wrongly is quite common in verbal communication between human beings.

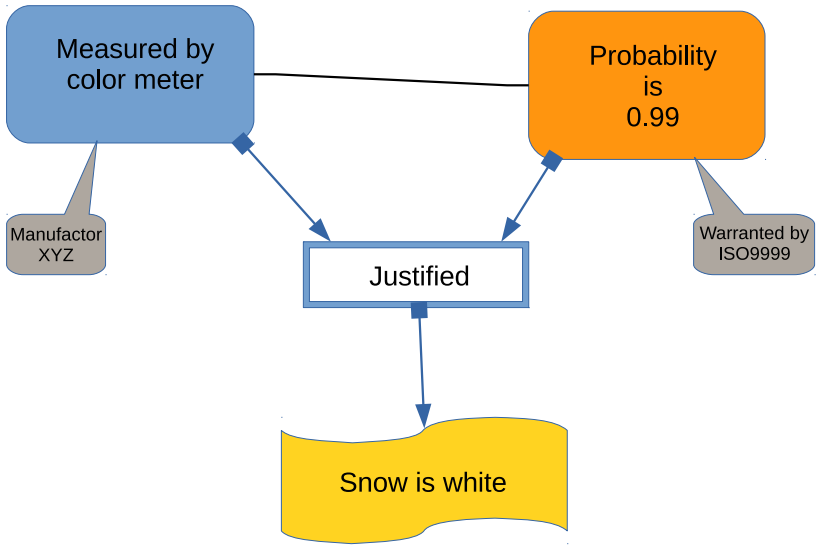

Figure 3.3: Justification.

claim that the reliability^p of a belief^p–forming process for justification is the proper approach.

The question *How is justification expressed?* is an interesting one, as whenever there is a need to evaluate in the context of ISA$_{bdi}$ whether or not there is justification for a belief^p, there must be some kind of representation of justification.[44] First, we argue that just like truth[45] justification should be expressed by a predicate, for example, *'is justified'* or *'Ju'*. The predicate can be expressed in a metadata describing the world where the proposition (for example, *Snow is white*) is stated. An example using RDF language (XML representation) is as follows[46]:

<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:ma="http://www.example.org/Materials/">

[44] An evaluator can be either ISA$_{bdi}$ itself, another ISA$_{bdi}$, or human being.
[45] See Section 3.3.
[46] This example is a hypothetical one.

```
<rdf:Description
    rdf:about="http://www.example.org/Materials/ColorOfMaterial/">
    <ma:material>snow</ma:material>
    <ma:color>white</ma:color>
    <ma:location>Ivalo</ma:location>
    <ma:justificationpredicate>isjustified</ma:justificationpredicate>
    <ma:justificationvalue>0.99</ma:justificationvalue>
    <ma:measurementmethod>coloranalyzer</ma:measurementmethod>
    <ma:certificating–institute>ISO</ma:certificating–institute>
    <ma:manufacturer>ABC</ma:manufacturer>
</rdf:Description>
</rdf:RDF>
```

Related to the second component *the sender of belief*^{p} S, we argue that in the context of ISA$_{bdi}$ S must have the justification for a belief^p in order to transmit it as a justified belief^p to a receiver R.[47] Otherwise S distributes incoherent information, for example, by implicitly implying that the belief^p is justified, even though it has no justification for it.[48] The case of *possible intermediate proxy distributor PD* is a different one. As *PD* does not add or decrease any value to the epistemic quality of a belief^p, it can be ignorant of the justification.

If we consider testimony as the method of distributing justified beliefs^p, then the sender S must have justification for hers/his/its beliefs^p. Now, the question is how this justification is transmitted to a receiver R? We can model the transmission of justified beliefs^p from ISA$_{bdi}S$ to ISA$_{bdi}R$ using the speech act theory [127]. We can compare ISA$_{bdi}$ performing an illocutionary act with a human being performing an illocutionary act, even though ISA$_{bdi}$ does not usually use any natural language to state the proposition, but an artificial language developed for ISA$_{bdi}$–to–ISA$_{bdi}$ information exchange. When we model ISA$_{bdi}$'s assertions using the speech act theory, we thus assume here that propositions represent the objects of justified beliefs^p. Speech acts, in which justified beliefs^p are typically transmitted, are those that ascribe justified beliefs^p by the use of a specific kind of sentence, a justified belief^p ascription.[49] Based on Searle's definition we can define the assert type of illocutionary act as follows [127]:

[47] Though the belief^p may not be justified in the context of the receiver R.

[48] Another way to distribute incoherent information is to implicitly imply that the belief^p is not justified, even though S has justification for it.

[49] In the context of communication between human beings this is like a speaker saying *I believe* instead of saying *I know*.

1. Propositional content: *Any justified belief p.*
2. Preparatory: *First, S has evidence (reasons, etc.) for the justification of p and second, it is not obvious to both S and R that R has justified belief p.*
3. Sincerity: *S has justification for p.*
4. Essential: *Counts as an undertaking to the effect that p represents a highly probable state of affairs.*

Based on this definition ISA_{bdi}'s assertion of justified belief^p can be specified as follows[50]:

1. Communicative act: *INFORM*
2. Message content: *Proposition*
3. Description: ***INFORM*** *indicates that agent S:*
   a) holds that the belief is justified,
   b) holds the evidence (e.g. reliability factor) of justification,
   c) intends that the receiving agent also concludes that the belief^p based on the proposition is a justified belief^p, and
   d) does not already believe that R has any knowledge of the justified belief^p.
   From the R's viewpoint, R believes that S has the justified belief^p based on the proposition, and after evaluation the evidence of justification R either considers the belief^p based on the proposition to be or not to be a justified belief^p.
4. Formal model: $<$ i, inform(j, ϕ, τ) $>$
   FP: $\text{JB}_i\ \phi \wedge \neg\ \text{JB}_i\ (\ \text{Bif}_j\ \phi \vee \text{Uif}_j\ \phi)$
   RE: $\text{JB}_j\ \phi \vee \text{B}_j\ \phi$
   Where:
   i = agent i; j = agent j;
   ϕ = proposition;
   τ = justification;
   FP = feasibility precondition; RE = rational effect;
   JB = justifiably believe; B = believe; Bif = believe if; and
   Uif = uncertain if.

We claim that testimony is not the actual source of justified belief^p, but just the transfer method. And in the contexts of ISA_{bdi} and DIDS justification itself shall be expressed along with belief^p.

[50] This example is modified from a definition made by FIPA [38].

3.5.7 Conclusion about Justified Belief in the context of ISA$_{bdi}$

We argue that an externalist theory of justification is the proper one in the context of ISA$_{bdi}$ because we cannot assume that it would be feasible or in some cases even possible to design and implement ISA$_{bdi}$ that would hold cognitively available internally all the factors needed for a belief^p to be epistemically justified.[51]

Process reliabilism provides the best theoretical (and also practical) foundations to evaluate justification in the context of ISA$_{bdi}$ because the justification status of a belief^p can be derived from the reliability^p of sensors, dependability of hardware equipment and software components, and observation of previous behaviour. The requirements of reliability^p can be derived from the dependability theory and requirements of computer science (application specific requirements). Reliabilism provides also truth–conduciveness.

We also support that justification is context–sensitive depending on the stake in question; the higher the stake is the higher the reliability^p requirement is for justification [63, 91]. This approach may lead to a situation where in one context a belief^p is justified but in another context there is no justification for the belief^p.

As there is always a possibility of a software error and hardware failure, we also support fallibilism in the context of ISA$_{bdi}$. In the case of testimony, we consider it to be just the transmission media of justified belief^p and justification itself [1].

We summarize our basic thoughts about justification in the contexts of ISA$_{bdi}$ and DIDS with the following definition (we call it pragmatic process reliabilism, PPR):

Definition. An epistemic agent's belief^{pc} *that* p is justified if and only if,

1. The epistemic agent believes p to be true;
2. The belief^{pc} was produced by sufficiently reliable^p processes $\mathbf{P}_i$; and
3. The required degree of reliability^p of the processes $\mathbf{P}_i$ is determined by the context where the epistemic agent uses his/hers/its belief^{pc} in reasoning and actions.

[51] Of course, this could be possible, if ISA$_{bdi}$ has only few justified beliefs^p.

3.6 Knowledge

In this section we discuss the concept of knowledge^p by exploring what is knowledge^p and what kind of entity is knowledge ^{p} in the context of ISA$_{bdi}$. We discuss various theories of knowledge^p, such as virtue epistemology, knowledge first, and reliabilism. We also discuss the role of testimony in the communication of knowledge^p.

There are several different kinds of knowledge^p: 1) propositional knowledge^p (*Snow is white.*), 2) knowledge^p of acquaintance (*Matti knows Maija.*), and 3) knowledge^p–how (*Matti knows how to ride a bicycle.*) [72, 99]. In the context of ISA$_{bdi}$ we deal with propositional knowledge^p, which has linguistic representations. Along with the development of AI, ISAs, and robotics the question can be raised, to what extent and benefits will and could knowledge^p–how and knowledge^p of acquaintance be transformed to propositional knowledge^p. Other representations, such as neural networks, are better solutions to represent knowledge^p of acquaintance and knowledge^p–how. For example, it might be quite feasible that in the future *A robot Biker knows how to ride a bicycle.*

There have been many attempts to analyse knowledge^p in the late 20^{th} Century, but so far there is no single, ubiquitously accepted definition. Therefore, there are also doubts whether knowledge^p is susceptible to analysis, at all [72]. We are of the opinion that if knowledge^p is susceptible to analysis, then it is also susceptible to an implementation based on AI. It is difficult to model and implement a concept without a proper analysis of the concept.[52]

In history philosophers have been of the opinion that knowing a thing involves, as well, knowledge of the limits of the thing: the limits that define both what the thing is and what it is not [40]. In the context of ISA$_{bdi}$ this requirement is interesting. As an example we can mention the scenario *"A pedestrian is crossing the road."* (see page 52) where, for example, ISA$_{bdi}$ must know whether the pedestrian–like feature nearby the road is a human being or a human–like statue.

When discussing knowledge^p in the context of ISA$_{bdi}$ we need to consider what it is for ISA$_{bdi}$ to know p. Is it just a list of conditions involving ISA$_{bdi}$ and p to hold or should it also comprise issues like the value of knowledge^p? Are there any benefits to ISA$_{bdi}$ having knowledge^p compared to ISA$_{bdi}$ having justified beliefs^p or mere beliefs^p? If there is, then how do we evaluate such things? We have already discussed the value of knowledge^p

[52] Timothy Williamson's approach *knowledge first* leads to a situation, in which other epistemic concepts are analysed on the basis of knowledge^p.

and the value problem of knowledge^p in the context of ISA$_{bdi}$ in Section 3.2.

First, we evaluate different knowledge^p theories and we also discuss the role of testimony in acquiring knowledge^p in the context of ISA$_{bdi}$. When we carry out the analysis of knowledge^p in the context of ISA$_{bdi}$, it is important to explore what are the main differences between ISA$_{bdi}$ to know something and a human being to know something? The epistemic literature comprises a huge number of analyses of knowledge^p in the context of human being; hence, when we are aware of the differences, we can adapt suitable ideas to the theory of knowledge^p for the contexts of ISA$_{bdi}$ and DIDS.

ISA$_{bdi}$'s main—most often only—purpose is to provide its users with services for which ISA$_{bdi}$ is designed and implemented. The requirements of the services determine what information is needed, and nothing else is inquired or inferred. Thus, the main reason to inquire, perceive, and infer information is to use it to provide the services, and ISA$_{bdi}$ has no information—belief^p, justified belief^p, and knowledge^p—for its own sake.[53] Based on this we argue that ISA$_{bdi}$ does not have intrinsic knowledge^p, but only instrumental knowledge^p to be utilized for practical reasons.

Knowledge^p as justified true belief^p (hereinafter JTB) is the base of almost all definitions of knowing [72]: Subject S knows *that p* if and only if

1. *p is true;*
2. *Subject S believes that p;* and
3. *Subject S is justified in believing that p.*

The truth condition (1) is generally accepted; it is intuitively plausible that false beliefs^p cannot be known. *Know* in the instance of the locution "x knows *that p*" is factive: if x knows *that p*, then *p* [98]. In the context of ISA$_{bdi}$ the truth condition raises the following question: If *know* is a factive verb—is knowledge that p knowledge of facts—then how does ISA$_{bdi}$ ensure the truthfulness of *that p*? For example, how does ISA$_{bdi}$ ensure that in our car driving example in Section 3.1 the belief^p established based on the perception of the pedestrian is true?

Jonathan Ichikawa and Matthias Steup discuss in their article *The Analysis of Knowledge* [72] that the truth of a proposition does not always require that anyone can know or prove that it is true. Not all truths are established truths. Truth is a metaphysical—not epistemological— notion:

[53] This will be the situation at least in the near future, as currently it is difficult to see how to design and implement an ISA$_{bdi}$ that would have some kind of intrinsic motive to value knowledge for the sake of knowledge itself.

truth is a matter of how things are, not how they can be shown to be. Knowledge is a kind of relationship with the truth—to know something is to have a certain kind of access to a fact. In the context of ISA_{bdi} this implies that it is not always required for ISA_{bdi} to prove the truthfulness of p.

There are three questions. First, what would be the right truth theory to be adopted? Second, what would be the certain kind of access to a fact? Third, which one is a better approach: infallibilism or fallibilism?

The first two questions are linked together, and we have discussed those above in Section 3.3. The reasoning of the proper truth theory results also in a certain kind of access to the fact. We have argued that the correspondence theory of truth is the most appropriate one in the context of ISA_{bdi}; therefore, a reliable correspondence as the world–to–world relation is the required kind of access to the fact.

Infallibilism and fallibilism are still open problems as Markus Lammenranta discusses in his article *We can't know* [81]. Fallibilism leads to paradoxes in some example cases, such as in the case of *the brain–in–a–vat*. And infallibilism leads to scepticism. Lammenranta settles this problem by taking into account the role of presuppositions. The sceptical hypothesis, for example *the brain–in–a–vat*, are ruled out in everyday contexts by the mutually accepted presuppositions. There is a distinction between what is said and what is meant by uttering a proposition. When knowledge^p is attributed to someone, we say that the evidence rules out all possibilities of errors (this is false), but what we mean or implicate is that the evidence rules out all the relevant possibilities of errors (this may be true). Now, the fallibilist explains ordinary uses of the term *know* by assuming that what is said and what is meant are both typically true. The infallibilist explains ordinary uses of the term *know* by assuming that what is said is false while what is meant is typically true. Infallibilism requires a failure–free implementation of ISA_{bdi} and its infrastructure. We claim that in the contexts of ISA_{bdi} and IDS infallibilism is not feasible because there is always a possibility of software and hardware failures. Therefore, there is always the possibility of p is not true even though it seems to be true.[54] Hence, infallibilism leads to scepticism in the context of ISA_{bdi}. Scepticism is not a good option, as the concept of knowledge^p is most likely beneficial in the context of belief^p exchange between human being and ISA_{bdi} as well as between ISA_{bdi}s. In addition, we are of the opinion that scepticism about ordinary knowledge is false. As pragmatists emphases that *"When we do go wrong, further discussion and investigation can identify and eliminate*

[54] We see that this is a similar kind of problem as Descartes' problem of certainty.

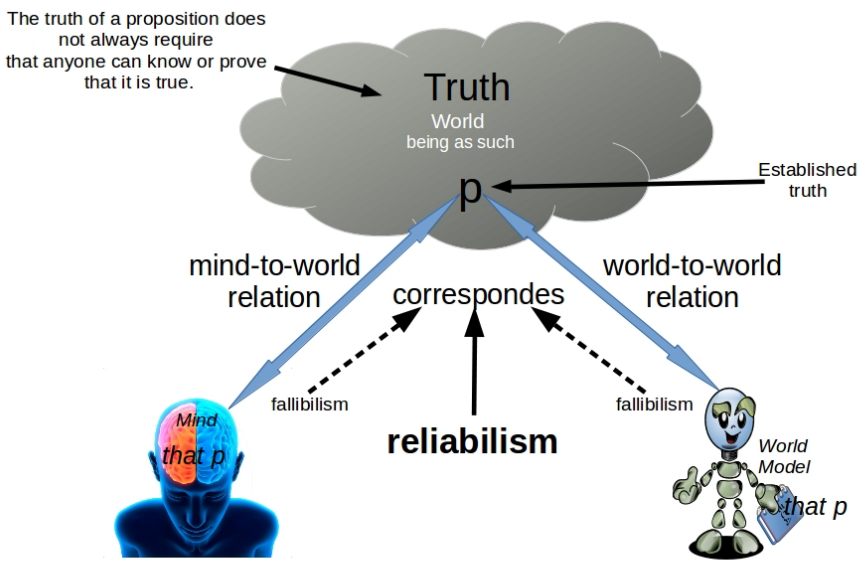

Figure 3.4: Truth condition.

errors, which is our best hope for escaping their damaging effects. ..... The focus of epistemological inquiry should not be on showing how we can possess absolute certainty, but on how we can develop self–correcting methods of inquiry that make fallible progress." [84].

We can summarize our approach to truth concerning knowledge^p in the context of ISA$_{bdi}$ as follows (illustrated in Figure 3.4):

1. The truth condition is essential in the definition of knowledge^p.
2. The concept of truth is metaphysical.
3. The concept of truth is not unambiguously defined.
4. The truth of a proposition does not always require that anyone can know or prove that it is true. Not all truths are established truths.
5. It is not always possible to verify the truth of a proposition in an infallible way, thus infallibilism leads to scepticism.
6. Scepticism is not an option.
7. Infallibility is not feasible.

Fallibilism denies that knowledge requires the impossibility of error. Therefore, unlike infallibilism fallibilism does not require absolutely failure–free implementations and services of ISA_{bdi}s. The history of computer science and computer–based services has clearly indicated that failure–free implementations and services are almost impossible to achieve; hence, we support fallibilism.

We have already discussed the belief^p condition (2) in Section 3.4 and the justification condition (3) in Section 3.3.

Since Gettier's article [52] *Is Justified True Belief Knowledge?*—luck should not have any role in achieving knowledge^p—several additional conditions have been proposed, such as the following ones [72]:

1. **No false lemmas**: an epistemic agent's belief^p *that p* is not inferred from any falsehood.

2. **Sensitivity**: an epistemic agent's belief^p *that p* is sensitive if and only if, if *p* were false, then epistemic agent would not believe *that p*.

3. **Safety^p**: if an epistemic agent were to believe *that p*, *p* would not be false, that is, *in all nearby worlds where an epistemic agent believes that p, p is not false*.

4. **Rule out all relevant alternatives**: an epistemic agent knows *that p* if only if the epistemic agent has ruled out relevant competing hypotheses to *p*.

5. **Anti–luck**: an epistemic agent's belief^p *that p* is not true merely by luck.

In the context of ISA_{bdi} every one of these conditions have problems. The *no false lemmas* condition is a strong requirement. It does actually require infallibilism when implementing it. And this requires, in turn, the failure free implementation of ISA_{bdi} that is utilized to create the lemma. We are of the opinion that this is not feasible. The *sensitivity* condition is difficult to model exactly for an implementation unless using probability because the concept *'if p were false'* is otherwise a vague term to be implemented in the contexts of ISA_{bdi} and DIDS. And if probability is used, then we argue that reliabilism is a more proper approach (see Section 3.6.5). The *safety^p* condition requires the implementations of the concepts of possible worlds and nearby worlds. But possible worlds and nearby worlds create complexity and are difficult to implement. This can, in turn, make it problematic to evaluate properly during runtime the epistemic quality of belief^p, for example, because of performance requirements and difficulty to

infer and instantiate real–time possible worlds. If the implementation of possible worlds and nearby worlds is required, then we claim that this shall be combined with reliabilism and the evaluation of possible consequences (see Section 3.6.5). The *ruling out all relevant alternatives* condition has, at least, two problems: how to decide what all alternatives are and which ones are not relevant (the concept of relevance is vague) [6], and performance—especially response time—can be severely affected when analysing relevant alternatives.[55] The *anti–luck* condition has, at least, two problems. The first problem is that luck is a vague term. Knowledge^p excludes luck, but luck comes in degrees. Then, how much luck does it take to be inconsistent with knowledge? And a distinction between different kinds of luck should be done [72]. Let us have an example: There is a reliable^p ISA$_{bdi}$, which predicts future occurrences. There are two cases. In the first case, a programmer of the ISA$_{bdi}$ has coded an algorithm correctly without properly understanding the algorithm; thus just being lucky. In the second case there is a fault in ISA$_{bdi}$ which causes an untrue proposition to be considered true, except in a certain case. The error causing the fault is unknown, thus the fault cannot be prevented. The possibility of the fault to occur is $10^{-1000000}$. But the proposition is true, if a certain proposition is used (probability 10^{-1000} in the fault case) in inferencing. Luck has a role in both these cases. Now, according to our intuition in the first case ISA$_{bdi}$ has knowledge^p based on the proposition, but not in the second case. The second problem is whether there is any possibility of ISA$_{bdi}$ to be aware of the role of luck when it evaluates the epistemic quality of the information based on the proposition.

3.6.1 Testimony about Knowledge

We have already discussed testimony in Section 3.5.6, where we discussed the basic ideas of testimony. In this section we discuss briefly the role of testimony in knowledge^p in the contexts of ISA$_{bdi}$ and DIDS. As in the case of justification we regard testimony as the transmission of knowledge^p from one epistemic agent to another one. The semantic character of testimony is important. It implies that without understanding the meaning of the symbols in which *that p* is asserted, an epistemic agent does not receive knowledge^p via testimony.

Robert Audi argues that testimony is essential for the spread of knowledge^p [10]. This is also the case in the context of DIDS. As already mentioned, Michael Dummet argues that *"Testimony should not be regarded*

[55] We are of the opinion that these kinds of solutions may violate the principle "keep it simple" a.k.a. Ockham's razor.

as a source, and still less as a ground, of knowledge^p: it is the transmission from one individual to another of knowledge^p acquired by whatever means." [35]. On the other hand, Robert Audi also states [10]: *"Although I know that p on the basis of your testimony only if you know that p, and I believe that p because you told me that p, your knowing that p is no more the (epistemic) basis of my knowledge^p than copper wire is the basis of electric current flowing through it to a light bulb. Your knowledge^p that p is required for successful transmission, but my knowledge^p is not based on your knowledge, if this entails more than its appropriately depending on it."* These contradictory statements raise the questions which one is the better approach in the contexts of ISA_{bdi} and DIDS, and what would be the requirements for testimony to transmit knowledge^p from one epistemic agent to another.

Jonathan Adler expresses default rules for testimony (anti–reductionism) in the following way [1]: *"If a sender S asserts that p to a receiver R, then, under normal conditions, it is correct for R to accept (acceptance as knowledge^p) S's assertion, unless R has a special reason to object."* And the knowledge^p norm of assertion states [1]: *"The sender S correctly asserts that p only if S knows (or represents oneself as knowing) that p."* Thus, according to the knowledge^p norm, if S does not know *that p*, S should not assert it as knowledge^p. For example, FIPA standards follow this anti–reductionist approach [39]. When there are one or more proxy epistemic agents in knowledge^p transmission, then, at least, the first epistemic agent in the chain of testimonial transfer must know *that p*; not every epistemic agent in the chain.

What is the difference between transmitting justified belief^p and transmitting knowledge^p in the testimonial transfer of information? First, what is the differentiating factor between knowledge^p and justified belief^p? As we support pragmatic process reliabilism, we claim that the reliability^p of a belief^p–forming process is the differentiating factor. The reliability^p (e.g. 0.999999) of the belief^p–forming process resulting in knowledge^p must be sufficiently higher than the reliability^p (e.g. 0.500000) of the belief^p–forming process resulting in mere justified belief^p.[56] Second, how does a receiver R differentiate whether it perceives justified belief^p or knowledge^p? A solution is that we add a reductionistic rule that requires a sender S to show the justifications for S's knowledge^p.[57] This can be done, for example, using

[56]Of course, there is a threshold between justified belief^p and knowledge^p, where the reliabilities^p close to each other. This creates problems (for example, the Sorites paradox); however, we believe that the problems can be resolved based on the application dependability requirements.

[57]This requirement is outside the original spirit of reliabilism.

metadata of knowledge^p (and justified belief^p, see Section 3.5). In addition, the transmission of knowledge^p must be dependable in a way that neither belief^p or the justification for belief^p is altered in any way.

A receiver R should always engage in some assessment of a sender S for trustworthiness. To believe what is asserted without doing so is to believe blindly, uncritically. The receiver R should always take a critical stance to the sender S, to assess it for trustworthiness. In the context of ISA$_{bdi}$ there are two possibilities. Either ISA$_{bdi}$ continuously evaluates the trustworthiness of the sender S, for example, using the methods presented by Sini Ruohomaa in [120] or having prima facie trustworthiness of the sender S based on the metadata describing justification—some kind of evidence, for example, the reliability^p of belief^p–forming processes—for knowledge^p.

Jennifer Lackey [80] argues that the absence of defeaters is a necessary condition for testimonial knowledge^p. The defeaters can be of the following types (S=sender, R=receiver):

1. *A defeater is a proposition which is believed by R to be true, yet indicates that R's belief^p that p is either false or unreliably formed or sustained.*

2. *A defeater is a proposition which R is justified in believing to be true, yet which indicates R's belief^p that p is either false or unreliably formed or sustained.*

3. *A defeater is a true proposition such that if the proposition was added to R's belief^p system, then R would no longer be justified in believing that p.*

What could the role of defeaters be in the context of ISA$_{bdi}$, as we consider that testimony is the transmission of knowledge^p, not the source of it. There can be new types of defeaters. For example, as we support pragmatic process reliabilism, a defeater could be a higher requirement of reliability^p for knowledge^p. The importance of consequences of R's action based on knowledge^p demands a higher reliability^p than the reliability^p of S's belief^p–forming process.

Sanford Goldberg [54] argues that the presence of an external defeater–detection system is critical for testimonially–based knowledge^p in the context of handicap receivers, such as children. We can consider ISA$_{bdi}$ to be a handicap receiver in the same sense and the same argument to be applied to testimonially–transmitted knowledge^p. In the context of ISA$_{bdi}$ there is an obvious requirement of a knowledge^p warranting/certifying service; especially when information comes from human beings.

3.6.2 Causal Theory about Knowledge

The main idea behind the causal theory of knowledge^p is to require a causal connection between a belief^p and a fact believed. A simple definition is as follows [72]:

Subject S knows *that p* if and only if

1. *p is true;*
2. *Subject S believes that p;*
3. *Subject S's belief^p that p is caused by the fact that p.*

According to this definition knowledge^p does not demand justification; instead, a causal connection between the belief^p and the fact believed. In the context of ISA$_{bdi}$ there is the question of what kinds of connections represent the right kinds of causality and are adequate. As the long history of discussions, numerous articles, and books about causality clearly advice, the issues related to causality are difficult, and the concept of causal connection is not well enough explicated to be generally implemented. And how to find out whether it is really a fact that caused the belief^p *that p*. We may ask, for example, in the environment of augmented reality, what is a fact. Therefore, we consider that the causal theory of knowledge^p falls short in the context of ISA$_{bdi}$.

3.6.3 Virtue Epistemology about Knowledge

Virtue epistemology defines that *S knows that p only if S acquires her belief in p by exercising some epistemic virtue and furthermore that a person who knows can be credited for her true belief in a way in which a person who has a mere true belief cannot* [147]. In other words, knowledge is true belief^p out of intellectual virtue [37].

In the context of ISA$_{bdi}$ the concept of intellectual virtue is definitely too vague to be modelled and implemented in general; unless we consider reliability^p to be the intellectual virtue. It is not clear what does intellectual virtue actually mean? There are several examples, such as Linda Zagzebski's definition [147, 168]: *"An act of intellectual virtue A is an act that arises from the motivational component of A, is something a person with virtue A would (probably) do in the circumstances, is successful in achieving the end of the A motivation, and is such that the agent acquires a true belief (cognitive contact with reality) through these features of the act. Knowledge is a state of true belief through these features of act."* As

we can see, this definition is not clear—actually quite far from being accurate enough—to be successfully modelled and implemented in the contexts of ISA_{bdi} and DIDS. This definitions would likely lead to a situation that there were many different kinds of implementations of the theory causing contradictory results of determining what is knowledge^p.

Ernest Sosa defines intellectual virtue using a AAA structure: accuracy, adroitness, and aptness [140]. He counts beliefs^p as performances, and therefore, beliefs^p can also be evaluated using the AAA structure. According to him *"we can distinguish between a belief's accuracy, i.e., its truth; its adroitness, i.e., its manifesting epistemic virtue or competence; and its aptness, i.e., its being true because competent"* [140]. Even though this definition is more accurate approach to knowledge^p, it leaves quite many open problems to be solved when modelling and implementing it. The problem of accuracy could be approached using the correspondes theory of truth. The problem of adroitness is more difficult because the concepts of epistemic virtue and competence are still vague. We could consider the correctness of an algorithm and its implementation to be epistemic virtue. This could be evaluated using the dependability measurements of computer science. The problem of aptness could be approached using the dependability attribute reliability^c of computer science, which defines that a correct service is not interrupted in any way. Still, Ernest Sosa's virtue epistemology would require a totally new evaluating and categorising system of dependability for ISA_{bdi}s and DIDS to be developed and implemented, which would be a huge and challenging task.

Therefore, we argue that virtue epistemology falls short of being the proper one in the contexts of ISA_{bdi} and DIDS.

3.6.4 Knowledge First about Knowledge

Due to the unsolved problems of the analysis of knowledge^p, Timothy Williamson has developed an approach to epistemology in which the notion of knowledge^p is explanatorily fundamental. He argues that knowledge^p cannot be analysed as a combination of truth, belief^p, and justification [164]. Thus, he turns the whole scheme of epistemology upside down. Belief^p and justification are analysed on the basis of knowledge^p.

We are of the opinion that this approach is very interesting, but it is not yet mature enough even in traditional epistemology in order to be considered as the knowledge theory for $ISAs^{bdi}$ and DIDS.

3.6.5 Reliabilism about Knowledge

We have already examined the basic ideas of reliabilism when we discussed justification in Section 3.5.5. We considered pragmatic process reliabilism to be the proper theory of justification in the context of ISA_{bdi} [63, 91]. The central idea of knowledge^p according to reliabilism is a reliable^p connection between the source and belief^p [55, 57]. A simple form of reliabilism about knowledge^p is defined as follows [72]:

Subject S knows *that p* if and only if

1. *p* is true;
2. S beliefs *that p*; and
3. S's belief *that p* was produced by a reliable cognitive process.

According to this definition knowledge^p does not require justification, but just a reliable^p cognitive process. What kind of process is considered to be a reliable^p cognitive process in the context of ISA_{bdi}? There are two issues: reliability^p and cognitivity. We can explicate reliability^p issues with the help of the dependability theories of computer science and warrant/certification services in the case of human beings. However, cognitivity cannot be explicated in the same way. We can say that ISA_{bdi} itself may execute cognitive processes. But what about a sophisticated thermometer, which informs reliably^p ISA_{bdi} about the temperature? Does it fulfil the requirements needed for knowledge^p when executing a 'cognitive' process? Thus, the requirement of processes in the forming of belief^p to be cognitive enough is difficult to explicate precisely, and therefore, we argue that this definition falls short in the context of ISA_{bdi}.

The cognitivity requirement of the belief^p–forming process is put aside in the definition of process reliabilism, and a fourth condition is added. Process reliabilism defines knowing as follows [57]:

Subject S knows that p if and only if

1. *p is true;*
2. *S believes p to be true;*
3. *S's belief that p was produced through a reliable process;* and
4. *A suitable anti–Gettier clause is satisfied.*

But as discussed above on the page 103 many of these anti–Gettier conditions create complexity that causes them to be problematic in the context of ISA_{bdi}. This raises the question which one of the anti–Gettier conditions is least problematic in the context of ISA_{bdi}. We are of the opinion that the *safety*^{p} condition combined with pragmatic encroachment is the right approach. The pragmatic encroachment can be stated in the following way: *A difference in pragmatic circumstances can constitute a difference in knowledge*^{p} [63, 72]. The basic idea is the following one: the reliability^p of a process together with the pragmatic importance of the consequences of belief^p determines whether an epistemic agent knows. An implementation of both safety^p and pragmatic encroachment can be done using the concept of possible worlds.

Pragmatic Process Reliabilism (hereinafter PPR) defines knowing in the contexts of ISA_{bdi} and DIDS as follows:

Definition. An epistemic agent knows that p if and only if

1. p is true;
2. The epistemic agent believes p to be true;
3. If the epistemic agent were to believe *that* p, p would not be false;
4. The epistemic agent's belief^{pc} *that* p was produced through reliable^p processes P_i; *and*
5. The reliability^p of the processes P_i either exceeds or is equal to the reliability^p requirements of the actions,
   (a) where the epistemic agent utilizes the belief^{pc} *that* p *and*
   (b) which are set by the expected consequences of the actions.

We have already discussed the first and second conditions above on the page 100. The third condition excludes beliefs^p, which are just lucky guesses—either through inferring or perception, from knowledge^p. The fourth condition includes the argument that the justification for belief^p is achieved through the reliability^p of belief^p–forming processes. The fifth condition brings into consideration possible factors created by the circumstances of an epistemic agent; especially, the pragmatic importance

of knowledge^p and the environment of an epistemic agent.[58] Knowledge^p is sensitive to the context.[59]

Let us evaluate PPR using our scenario in Section 2.1.1. As already discussed, there are three different propositions:

1. P1: *The correct diagnosis is lateral malleolus fracture.*
2. P2: *The correct diagnosis is bimalleolus fracture.*
3. P3: *The correct diagnosis is trimalleolus fracture.*

Our intuition says that the belief, the object of which is proposition P1, is not knowledge. Now, according to PPR: *An epistemic agent knows that p if and only if*

1. *p is true;*
   P1 is not true; therefore, it would be straightforward to claim that PPR fulfils directly our intuition. But there is a viewpoint that deserves deeper consideration. The problem is following: Can an attributer actually determine that whether there is such a reliable^p correspondes as the mind (world)–to–world connection that can be considered to be the demanded truth–conduciveness of the belief^p–forming process in order to decide whether p is true or not. But as we support fallibilism, and we see that the role of truth is at the background scene, we cannot argue that this is the dominant factor in the determination whether PPR is according to our intuition.
2. *The epistemic agent believes p to be true;*
   The physician at the district hospital believes P1 to be true, but other epistemic agents do not believe P1 to be true (except Matti during the first phase). Hence, the context of the attributor affects the evaluation of this requirement, and PPR either fulfils or does not fulfil our intuition.
3. *If the epistemic agent were to believe that p, p would not be false;*
   P1 is false in the actual world; therefore, excluding the role of luck is not relevant in this case.
4. *The epistemic agent's belief that p was produced through reliable^p processes $\boldsymbol{P}_i$;*
   We are of the opinion that the imaging process exercised at the district hospital is reliable^p in general.

[58] In computer science this is called context–awareness.

[59] Hereinafter belief, justified belief, and knowledge without superscripts ^{p} and ^{c} refer these terms in the contexts of both human being and ISA$_{bdi}$.

5. *The reliability^p of the processes $\boldsymbol{P}_i$ either exceeds or is equal to the reliability^p requirements of the actions*,

   (a) *where the epistemic agent utilizes the belief^p that p*
   The belief is utilized to provide good medical care for Matti.

   (b) *which are set by the expected consequences of the actions.*
   The expected consequences include full recovery of Matti's ankle and smallest possible healthcare costs. Therefore, there is a high requirement to avoid all the errors that would cause any failures in the medical care, and this, in turn, demands very high reliability^p.

   Despite of being reliable^p in general, the imaging process does not fulfil the reliability^p requirements that are high because of the potential consequences of a failure.

We can conclude that according to PPR P1 is not knowledge. PPR is according to our intuition.

Proposition P2 is more complicated, and it demands a more thorough evaluation. Our preliminary intuition says that the belief, the object of which is proposition P2, is knowledge, but after more careful thinking about possible consequences of failure our intuition says that it is not knowledge. According to PPR:

The epistemic agent knows that p if and only if

1. *p is true;*
   P2 is not true. But, again, the same applies as in the case of P1.

2. *The epistemic agent believes p to be true;*
   The CADx system as well as the physician at EMA, Matti, and TMA application believe P2 to be true.

3. *If the epistemic agent were to believe that p, p would not be false;*
   P2 is false in the actual world, therefore, excluding the role of luck is not relevant in this case.

4. *The epistemic agent's belief that p was produced through reliable^p processes $\boldsymbol{P}_i$;*
   The imaging process exercised at the Finnish radiology center is reliable^p.

5. *The reliability^p of the process $\boldsymbol{P}_i$ either exceeds or is equal to the reliability^p requirements of the actions*,

(a) *where the epistemic agent utilizes the belief that p*;
The belief is utilized to provide the best possible medical care for Matti and to reduce possible post–trauma costs.

(b) *which are set by the expected consequences of the actions.*
After all, the imaging process does not fulfil the high reliability^p requirements, which are set by the demand for Matti's full recovery from the trauma. The consequences of partial recovering would require both Matti to change his occupation and the insurance company to pay Matti compensations, which would otherwise be unnecessary.

We conclude that according to PPR that the information based on P2 is not knowledge. PPR is according to our intuition.

Our intuition says that the belief, the object of which is proposition P3, is knowledge. P3 is based on the result of the highly reliable^p process, and there exist no defeaters. According to PPR:

An epistemic agent knows that p if and only if

1. *p is true;*
P3 is true. At least, it is based on the best available empirical practice, and therefore, we argue there is such a reliable^p correspondes as the mind (world)–to–world connection that can be considered to be the demanded truth–conduciveness of the belief–forming process in order to decide whether p is true or not. But, again, as we support fallibilism in this kind of empirical contexts, there is a possibility that P3 is untrue. However, we see that the role of truth is at the background scene in this kind of empirical contexts. Again, we cannot argue that this is the dominant factor in this case.

2. *The epistemic agent believes p to be true;*
All the epistemic agents involved believe P3 to be true.

3. *If the epistemic agent were to believe that p, p would not be false;*
P3 is true in the actual world, therefore, excluding the role of luck is relevant in this case. The question is which are the nearby worlds, that need to be taken into account. We can consider the following possible worlds:

   - *Another orthopaedist at Helsinki University Hospital could have operated on the ankle.*

We consider this case to be a nearby world. Another orthopaedist would have come to the same conclusion. There is no luck involved.

- *Matti could have been transferred to another Finnish university hospital.*
  We consider this case to be a nearby world. Another orthopaedist at a different university hospital would have come to the same conclusion. There is no luck involved.
- *Matti could have been operated in a Finnish district hospital.*
  We consider this case to be a nearby world. Another orthopaedist at a district hospital would have come to the same conclusion, because of the high medical level of Finnish district hospitals. There is no luck involved.
- *The operation of the ankle could have failed.*
  We consider that this case is not a nearby world because the failure of an ankle operation is very rare.
- *Matti could have been operated on in a local hospital in Thailand.*
  We consider that this case is not a nearby world because Matti was capable to travel back to Finland for the operation; therefore, Finnish travellers are normally transferred back to Finland.
- *Matti could not have been operated, at all.*
  We consider that this case is not a nearby world because it would have been erroneous medical care.

4. *The epistemic agent's belief that p was produced through a reliable^p process $\boldsymbol{P}_i$;*
   The operation, which was carried out at Helsinki University Hospital, can be considered to fulfil all the requirements of a reliable^p process.

5. *The reliability^p of the processes $\boldsymbol{P}_i$ either exceeds or is equal to the reliability^p requirements of the actions*,

   (a) *where the epistemic agent utilizes the belief that p*
   P3 is utilized to provide the best possible medical care for Matti and to reduce possible post–trauma costs.

   (b) *which are set by the expected consequences of the actions.*
   The operation process does fulfil the high reliability^p requirements, which are set by the demand of Matti's full recovery of the trauma. The consequences of partial recovering would require both Matti to change his occupation and the insurance

company to pay Matti compensations, which would otherwise be unnecessary.

We conclude that according to PPR P3 is knowledge. PPR is according to our intuition.

PPR begs some questions in the context of ISA_{bdi}. The first one deals with planning and implementation of ISA_{bdi}. What is actually required to plan and implement ISA_{bdi} fulfilling the requirements of PPR? The second question deals with the runtime environment of ISA_{bdi}. How can ISA_{bdi} be aware of expected (or potential) consequences of its actions? We discuss these issues in more detail in Section 6.2.

3.6.6 Conclusion about Knowledge in the context of ISA_{bdi}

We argued that JTB added with any of the anti–Gettier conditions (sensitivity, safety^p, or rule out all relevant alternatives) is not suitable in the context of ISA_{bdi}. The same applies also to the causal theory of knowledge, virtue epistemology, and knowledge first.

Process reliabilism provides the best alternative when the fifth condition (pragmatic encroachment) is added. PPR defines knowing as follows:

An epistemic agent knows that p if and only if

1. *p is true;*
2. *The epistemic agent believes p to be true;*
3. *If the epistemic agent were to believe that p, p would not be false;*
4. *The epistemic agent's belief that p was produced through reliable processes P_i ;* and
5. *The reliability^p of the processes P_i either exceeds or is equal to the reliability^p requirements of the actions*,
   (a) *where the epistemic agent utilizes the belief that p* and
   (b) *which are set by the expected consequences of the actions.*

The fifth condition begs, at least, two questions in the context of ISA_{bdi}. First, what is required to plan and implement ISA_{bdi}? Second, how can ISA_{bdi} at runtime be aware of expected (or potential) consequences of its actions? We give some answers to these questions in Section 6.2.

In the context of ISA_{bdi} there exists only instrumental value of knowledge, and the question of knowledge value concerns the value of knowledge

over the value of justified belief; not the value of knowledge over the value of true belief.

3.7 Trust

In this section we discuss the roles of trust and trustworthiness in the contexts of ISA_{bdi} and DIDS. We also enhance the explication of trust.

Trust is important because it is an essential attitude when forming a relationship with others. We hope that others are trustworthy. Thus, trust comprises an attitude that there is a risk that others will deceive us. In the contexts of ISA_{bdi} and DIDS trust has a role in testimony. According to Judith Baker there are three kinds of trust [13]. In the first kind, people trust other people because they cannot check all the bases for establishing a belief. The formed belief is not resistant to counter–evidence. The second kind of trust involves more than people's willingness to accept other people or to assume things on trust. People may judge an individual or a thing on the basis of a non–ordinary belief. For example, a person might think a salesman is honest because he just looks honest. Once again, the formed belief is not resistant to counter–evidence. A common feature among these two kinds of trusts is that they do not directly rely on evidence. The third kind of trust is the case in which people think it is rational to hold a belief even though there is a counter–evidence. This kind of trust can be called friendship trust.

Trust and trustworthiness are important because they are fundamental concepts of relationships. But these terms are not easy to grasp. We can approach them from several viewpoints: philosophy, computer science, psychology, sociology, religion, etc. We will explore briefly the following dimensions of trust:

1. The conceptual nature of trust and trustworthiness

2. The epistemology of trust.

For example, one of the definitions of *trustworthiness* in the context of human beings is as follows: *Trustworthiness is a moral value considered to be a virtue* [28, 93]. Thus, a trustworthy person is someone in whom you can place your trust and rest assured that the trust shall not be betrayed. As we can see in the discussion below, this definition is far from being an accurate one. And in the context of ISA_{bdi} moral values and virtues are out

of scope, as we cannot argue yet that ISA_{bdi} could have such properties.[60] However, there are other philosophical dimensions of trustworthiness that are fruitful to explore in the context of ISA_{bdi}. Trustworthiness and trust need to be discussed together because they are strongly linked to each other [65].

Trusting [93] requires that a trustor can tolerate some level of risk or vulnerability to the failure by a trustee to do or to be what the trustor depends on in the trustee. Hence, trusting requires that a trustor can 1) be vulnerable to the trustee; 2) think well of the other, at least, in certain domains; and 3) be optimistic that the trustee is competent in certain respects. For trust to be warranted (well–grounded), both parties must be trustworthy.

In the context of computer science the definition is as follows: *The trustworthiness of a component is defined by how well it secures a set of functional and non–functional properties, deriving from its architecture, construction, and environment, and as evaluated as appropriate.* This definition provides us with a better starting point in the context of ISA_{bdi}. A more precise definition comes from the Committee on Information Systems Trustworthiness [125]: *Trustworthiness of distributed systems asserts that the system does what is required despite environmental disruption, human user and operator errors, and attacks by hostile parties and that it does not do other things. Design and implementation errors must be avoided, eliminated, or somehow tolerated. Addressing only some aspects of the problem is not sufficient. Moreover, achieving trustworthiness requires more than just assembling components that are themselves trustworthy.*

In philosophy the terms trustworthiness and trust are usually used in the context of the relationship between human beings (case 0 in Figure 3.5). Trust is generally a three–party relation: A trusts B to do X [66]. But when discussing those terms in the contexts of DIDS and ISA_{bdi}, the matter is not clear. Ori Freiman discusses some of these problems in the context of Internet of Things in his article *Towards the Epistemology of the Internet of Things* [46]. Do we have any grounds to extend the relationship to include artificial intelligent agents and components. For example, how does the trust–trustworthiness scheme change, when instead of stating that 'Juhani trusts Maria to inform him of the correct water temperature.', we state:

– 'Juhani trusts ISA_{bdi} to inform him of the correct water temperature.'

[60] Though, for example, Nick Bostrom argues in his book *Superintelligence: Paths, Dangers, Strategies* [21] that in the future these features are feasible.

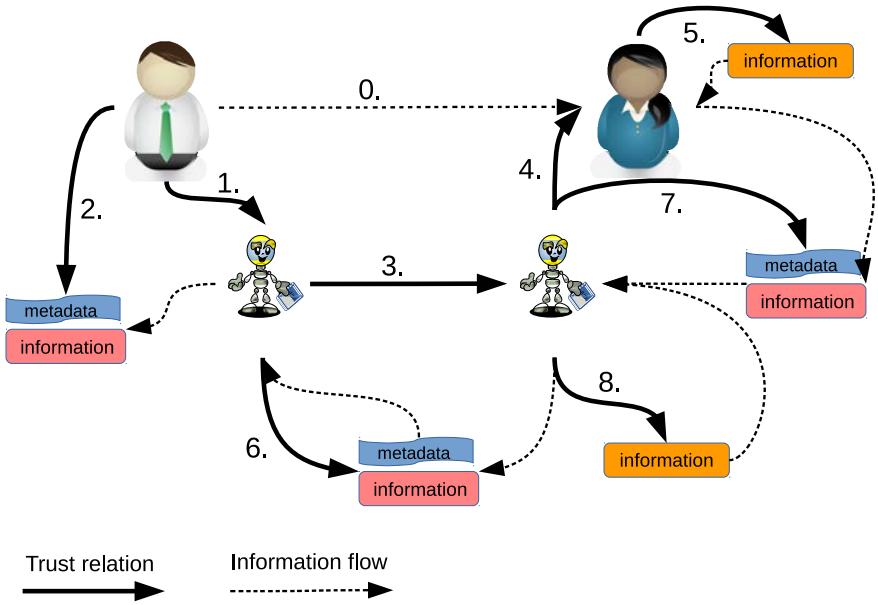

Figure 3.5: Cases of trust.

– 'Juhani trusts a justified belief of the correct water temperature.'
– 'ISA$_{bdi}$ trusts Maria to inform it of the correct water temperature.'
– 'ISA$_{bdi}$ trusts another ISA$_{bdi}$ to inform it of the correct water temperature.'

In the environment of DIDS we can point out several different cases, where the trust–trustworthiness scheme plays a role in addition to trust between human beings. The cases are as illustrated in Figure 3.5:

1. Human being trusts ISA$_{bdi}$.
2. Human being trusts a piece of information (e.g. knowledge or justified belief) provided by ISA$_{bdi}$.
3. ISA$_{bdi}$ trusts another ISA$_{bdi}$.
4. ISA$_{bdi}$ trusts human being.
5. Human being trusts a piece of information (obtained, for example, from some sort of a sensor).

6. ISA_{bdi} trusts a piece of information (e.g. knowledge or justified belief) provided by ISA_{bdi}.

7. ISA_{bdi} trusts a piece of information (e.g. knowledge or justified belief) provided by human being.

8. ISA_{bdi} trusts a piece of information (obtained, for example, from some sort of a sensor).

We argue that the first case *"human being trusts* ISA_{bdi}*"* is similar to the case of *"human being trusts human being"*. The above–mentioned three requirements are fulfilled. First, human being is vulnerable to ISA_{bdi}. Second, human being usually thinks well of ISA_{bdi}, the services of which she/he uses at least in a specific domain. And third, human being is optimistic that ISA_{bdi} is competent in required respects.

The second case is interesting, for example, in a case where human being has no trust–related information about ISA_{bdi}, which is the source of the piece of information. What does *trust a piece of information* actually mean? As mentioned above *trust is generally a three–party relation: A trusts B to do X*. We can enhance this in the following way: A trusts B to have a property X. We can explicate it in the following way: Human being has strong confidence that the piece of information has a certain property, which in the case of a proposition is truthfulness and/or justifiedness. Now, if the proposition does not actually have the required property, then we can say that the proposition has betrayed human being. The above–mentioned three requirements are fulfilled in this case, as well. First, human being is vulnerable to the truthfulness or justifiedness of the proposition because a false proposition can cause actions, which might be harmful to human being. Second, human being usually thinks well of the proposition, which has either instrumental or intrinsic value. And third, human being is optimistic that the proposition can be the object of justified belief and knowledge.

We see that the third case ISA_{bdi} *trusts another* ISA_{bdi} is similar to the first one. First, ISA_{bdi} is vulnerable to another ISA_{bdi}. Second, it is quite feasible to consider that ISA_{bdi} thinks well of another ISA_{bdi}s, which services she/he uses at least in a specific domain. And third, ISA_{bdi} has a positive belief that ISA_{bdi} is competent in certain respects.

The fourth case ISA_{bdi} *trusts human being* is similar to the case of *human being trusts human being*. The above–mentioned three requirements are fulfilled, as well. First, ISA_{bdi} is vulnerable to human being. For example, if human being provides ISA_{bdi} with an unjustified or false proposition, then ISA_{bdi} might not provide a reliable^p (dependable) service. Second, it is quite feasible to consider that ISA_{bdi} thinks well of human

being at least in a specific domain. And third, ISA_{bdi} has a positive belief that human being is competent in certain respects. The cases from five to eight similar to cases from one to four.

Trust can be warranted in the sense of being plausible. The term 'warranted' comprises concepts 'justified', 'well–grounded', and 'plausible'. A piece of information is trustworthy meaning that trust is warranted, when it is for example [93]:

1. *"Plausible, only if the conditions required for trust exist. Knowing what these conditions are requires understanding the nature of trust."*

2. *"Plausible, only when it is possible for one to develop trust, given one's circumstances and the sort of mental attitude trust is. For example, trust may not be the sort of attitude that one will oneself to have without any evidence of a person's trustworthiness."*

3. *"Well–grounded, only if the trustee is trustworthy, which makes the nature of trustworthiness important in determining when trust is warranted."*

4. *"Justified, sometimes when the trustee is not in fact trustworthy, which suggests that the epistemology of trust is relevant."*

5. *"Justified, often because some value will emerge from the trust or because it is valuable in and of itself. Thus, the value of trust is important."*

We summarize our thoughts about trust and trustworthiness by stating that in the contexts of ISA_{bdi} and DIDS they are important because they are fundamental factors in the co–operation between ISA_{bdi}s and between ISA_{bdi} and human being. Without trust and trustworthiness there hardly would exist any co–operation. In the context of ISA_{bdi} the concepts *trust* and *trustworthiness* must cover also cases that are more sociological than technological by nature. We also extended trust to cover the attitude towards entities such as propositions by stating *A trusts B to have a property X*, where B is a proposition. Proposition B is trustworthy if it actually has the property X, otherwise it betrays A, if A has established a strong confidence based on proper factors that B has the property X. Trust and trustworthiness are essential background components in the evaluation of the dependability of ISA_{bdi} and IDS.

3.8 Possible Objections

In this section we discuss some potential objections to our approach. We consider that the following four ones are the most important objections: anthropomorphism, joint epistemic theories, pragmatic process reliabilism as joint epistemic theory, and implementability.

Objection 1: Anthropomorphism

Even though the discussion created by John R. Searle on the capability of an artificial entity (a digital computer) having intentionality, understanding, and qualia has faded during 2010s, there still are doubts about AI–based entities to be capable of having semantic content of the sort that is essential to human cognition. In the context of this thesis the question is whether ISA$_{bdi}$s (also robots) can have belief, justified belief, and knowledge. The fundamental claim is as follows [129]: *"The purely formal or abstract or syntactical processes of the implemented computer program could not by themselves be sufficient to guarantee the presence of mental content or semantic content of the sort that is essential to human cognition. Of course a system might have semantic content for some other reason, but it does not apply to Strong Artificial Intelligence, any more. The basic structure of the Chinese Room Argument is rather obvious: the distinction between syntax and semantics and the distinction between simulation and duplication."*

We can answer the objection using the following thought experiment: We have two exactly similar rooms, which have equipment to carry out the following training session. On the floors there are several tools, such as a screwdriver, a hammer, a saw, etc. In one room there is a child to whom a man teaches how to use the tools, and in another room there is a robot[61] to whom another man teaches how to use the tools. The teaching session goes in the following way: The teacher says to the child/robot *"Bring me the hammer."* As, the first time, the child/robot does not know which tool is the hammer, she/it picks up an arbitrary object which happens to be the screwdriver, and brings it to the teacher. The teacher says in both cases: "No, this is not the *hammer*; this is the *screwdriver*." The child/robot takes the screwdriver back to its place. Next the child/robot takes the hammer and brings it to the teacher. The teacher says in both cases: "Yes, this is the *hammer*, thank you". Next the teacher says to the child/robot *"Bring me the screwdriver."* Now, the child/robot does know, which tool is the

[61] The robot has required visual tools to recognize various objects, and limbs to move and catch objects.

screwdriver, she/it brings the screwdriver. The teacher says in both cases: "Thank you".[62]

Now, according to Searle, the child (mind) has a semantic mental content; therefore, the child understands and is able to reason. But the robot does not have a semantic mental content; hence, the robot does not understand (however, the robot is able to reason). The robot only simulates[63] understanding of the meanings of the symbols, and there is not the right kind of causal relationship[64] [130]. Our intuition says that in both cases there is semantics, thus understanding, involved. Our defence can be summarized in the following claims: 1) Externalist attitude: states of a physical entity get their content through causal connections to the external reality they represent. This is not limited only to human beings. 2) ISA_{bdi} can have propositional attitudes if it has the right causal connections to the world. 3) The syntactically specifiable objects over which computations are defined can possess semantics; it is just that the semantics may not be involved in the specifications. For example, in the context of ISA_{bdi} we can have semantics involved using metadata about information. 4) Programming is precisely what could give something a mind. Therefore, we argue that ISA_{bdi} do have belief, justified belief, and knowledge.

Objection 2: Joint Epistemic Theories

The following question can be raised: Is there real need for joint epistemological theories? This question is valid because the worlds (operating environments) of human beings and ISA_{bdi}s can be considered to be quite different from each other. And so far human beings have generally seen computer–based entities not to be independent, but operate as useful tools managed by human beings. Therefore, there is no need for epistemic theories for computer–based entities.

The environment of computer–based systems (intelligent distributed systems and intelligent software agents) is changing rapidly due to the current strong development of AI and robotics. There will be independent ISA_{bdi}s, robots, and other kinds of ISAs (e.g. based on deep learning). They will offer services to human beings, so that human beings do not know that the services are provided by independent computer–based entities, which

[62] This thought scenario can be extended by adding a teaching session discussing what can be done with *hammer* and *screwdriver*.

[63] We cannot prove it to be duplication because we don't know yet how a human brain actually produces understanding and what the limits of understanding (mind) are.

[64] According to Searle the causal relationship should be a bottom–up relationship and not an input–output relationship.

are capable to learn, change their behaviour, deny or decline to offer services, etc. Therefore, service users cannot make any difference between whether the services are provided by human beings or ISA$_{bdi}$s. And in the context of DIDS a service can be provided by co–operation between human beings and independent ISA$_{bdi}$s. Therefore, we argue that joint epistemic theories provide a better ground for the dependability of ISA$_{bdi}$s and DIDS. If there were a contradiction between human beings and ISA$_{bdi}$s about the epistemic theories, thus the epistemic quality of information, would it lead to undependable IDS. The epistemic quality of information affect the dependability of IDS, as discussed in Section 2.2.2.

Objection 3: Pragmatic Process Reliabilism as Joint Epistemic Theory

Reliabilism has been discussed including objections in several articles, such as [60, 57, 58, 148, 149]. For example, Jonathan Vogel in his article *Reliabilism Leveled* [148] points out that reliabilism has problems dealing with higher–level or reflective knowledge. Therefore, he argues that the inability of reliabilism to account for reflective knowledge has its roots in a more basic deficiency, and he considers that in general knowledge is not either a reliably true belief or a belief that results from a reliable^p process. He sees that knowledge is a kind of human success.

We argue that the above objection is not strong enough in the environment, where human beings and ISA$_{bdi}$s (or other intelligent artificial entity based on AI) work in co–operation to produce dependable services. First, there is no consensus among epistemologists on a general, overall valid epistemic theory of justification and knowledge, and therefore, most proper one for the joint environment shall be selected. Second, we have already argued (see Sections 3.5 and 3.6) that PPR is the most proper one to be the joint epistemic theories. Third, we argue that the problems of higher–level or reflective knowledge are not so decisive factors that it would overcome the benefits of PPR over other possible epistemic theories in the environment of human beings, DIDS, and ISA$_{bdi}$s. The actual, pragmatic role of a higher–level or reflective knowledge in producing, storing, and utilizing beliefs is not so significant that it would be more meaningful than the capability to evaluate efficiently the epistemic quality of information. The higher–level knowledge (to know that one knows p) is either very seldom needed or practical in the context of ISA$_{bdi}$.

Objection 4: Implementability

Epistemological discussions have been going on at least for two thousand years including a hidden idea that only a human mind can really operate according to a developed epistemic theory. Now, the scheme is being and will be changed by AI. When adopting a theory of justification and knowledge (we can call it as applied epistemology), we must take into account the implementability of the theory. We have already argued in Section 3.5, that process reliabilism faces less implementation problems than other theories in justifying ISA_{bdi}'s beliefs. PPR still faces implementation problems, especially in the case of knowledge, that have not yet been solved by AI. The problems deal with the evaluation of possible consequences of an action (various counterfactual worlds), which in turn affect the evaluation of the epistemic quality of belief. However, we are of the opinion, that those problems will be solved by the future development of AI.

3.9 Summary of Six Concepts

In the environment where human beings and ISA_{bdi}s work in co–operation to produce dependable services we need the same epistemological base for both parties.

Truth

We regard truth as an important concept in the co–operation between ISA_{bdi}s and human beings, but it plays its role in the background. It is intuitive to say that it cannot be known what is not true. We argued that truth is the substantive property (inflationist approach), and there exists the property F (correspondence) such that any proposition, if true, is so by virtue of being F and this is a fact that is not transparent in the concept of truth. So correspondence is necessary and sufficient for explaining the truth of any true proposition p. The reliabilist theories of knowledge scope truth well—correspondence of the world–to–world connection—in terms of the truth–conduciveness of a belief–forming process.

Trust and Trustworthiness

In the co–operation between ISA_{bdi}s and human beings trust and trustworthiness are important, as they are fundamental factors in the successful service provisions. Trust and trustworthiness cover situations that are more sociological than technological by nature. We enhanced trust to cover the

attitude towards entities such as propositions by stating *A trusts B to have a property X*, where B is a proposition. Proposition B is trustworthy if it actually has the property X, otherwise it betrays A, if A has established a strong confidence based on proper factors that B has the property X.

Summary of Definitions

Our basic understanding of the required epistemic base can be stated as follows:

Belief An epistemic agent has beliefs that have similar features to compared with traditional human being's beliefs. Belief is defined as follows:

Belief is a propositional attitude,

1. which is the state of having an opinion about something to be the case;
2. which is created by its actual and potential causal relations to sensory stimulations, behaviour, and/or other propositional attitudes; and
3. the representation of which—structured if necessary—is stored in a linguistic form.

Justified Belief Pragmatic process reliabilism explicates the justification in the joint context of ISA_{bdi} and human being. The definition of justification is as follows:

An epistemic agent has justification for her/his/its belief *that p* if,

1. The epistemic agent believes p to be true;
2. The belief was produced by sufficiently reliable^p processes $\mathbf{P}_i$; and
3. The required degree of reliability^p of the processes $\mathbf{P}_i$ is determined by the context where the epistemic agent uses his/hers/its belief in reasoning and actions.

Knowledge Pragmatic process reliabilism explicates well the way we understand the concept of propositional knowledge. Knowledge is defined as follows:
An epistemic agent knows *that p* if and only if

1. p is true;
2. The epistemic agent believes p to be true;
3. If the epistemic agent were to believe *that p, p* would not be false;
4. The epistemic agent's belief *that p* was produced through reliable^p processes $\mathbf{P}_i$; and
5. The reliability^p of the processes $\mathbf{P}_i$ either exceeds or is equal to the reliability^p requirements of the actions,
   (a) where the epistemic agent utilizes the belief *that p* and
   (b) which are set by the expected consequences of the actions.

In the context of ISA$_{bdi}$ there exists only an instrumental value of knowledge.

Conclusions of Six Concepts

We have reasoned that PPR as the justification theory and knowledge theory, and testimony as the transfer method of justified belief and knowledge are the most proper ones in the contexts of ISA$_{bdi}$ and DIDS.

We have also come to the conclusion that truth is a very complex concept, and it is not quite clear what its actual role in the context of ISA$_{bdi}$ is. We argued that truth is the substantive property (inflationist approach), and there exists the property F (correspondence) such that any proposition, if true, is so in virtue of being F and this a fact that is not transparent in the concept of truth. So correspondence is necessary and sufficient for explaining the truth of a true proposition p in the context of ISA$_{bdi}$. In the case of ISA$_{bdi}$ we assumed that the reliability^p—if it is high enough—of the belief–forming process, which is used to establish the correspondence, is adequate to indicate truthfulness. And in the context of ISA$_{bdi}$ truth plays its role in the background; though, in the context of various logics truth is the important factor.

In Section 3.5 we argued that justification is context–sensitive, and pragmatic process reliabilism is the proper theory of justification. We also argued in Section 3.6 that the pragmatic process reliabilism theory of knowledge is the most appropriate one in the context of ISA$_{bdi}$. We are of

the opinion that PPR explains truth adequately enough—correspondence as the world–to–world connection—in terms of the truth–conduciveness of ISA_{bdi}'s belief–forming process.

The epistemic contexts of ISA_{bdi} and DIDS can be very complicated consisting of many sources of information and many kinds of information. We discuss this issue in more detail in Chapter 4.

Chapter 4

Belief as Dependability Factor

In this chapter the main scope of our discussion is to explore various roles and meanings of knowledge, justified belief, and belief when taking into account dependability of information services provided by ISA_{bdi} and DIDS.

When a human being has a knowledge—or just justification for a belief—and she/he needs to act based on this belief then she/he usually trusts that acting based on this belief will lead to a successful result. Thus, belief is a dependability factor in human being's actions. We claim that this is the case also in the actions of ISA_{bdi}.

4.1 Justifiably be Trusted

The term trustworthiness in the context of computing systems usually means that systems are secure, available, and reliable[c]. Thus, a trustworthy system does what people (its developers and users) expect it to do and not something else despite environmental disruption, errors made by human users and operators, hardware failures, and attacks by hostile parties. Design and implementation errors must be avoided, eliminated, or somehow tolerated. It is not sufficient to address only some of these dimensions, nor is it sufficient simply to assemble components that are themselves trustworthy. Trustworthiness is holistic and multidimensional [125]. But we are of the opinion that this approach does not adequately address the epistemological concept required by ISA_{bdi} and DIDS.

As discussed in Section 2.1.2 a dependable system shall either deliver a service that can justifiably be trusted or be capable to avoid service failures that are more frequent or more severe than is acceptable to the users. We also noted that there is a causal relationship between these two options. Justified belief and knowledge are qualitative factors, but the evaluation

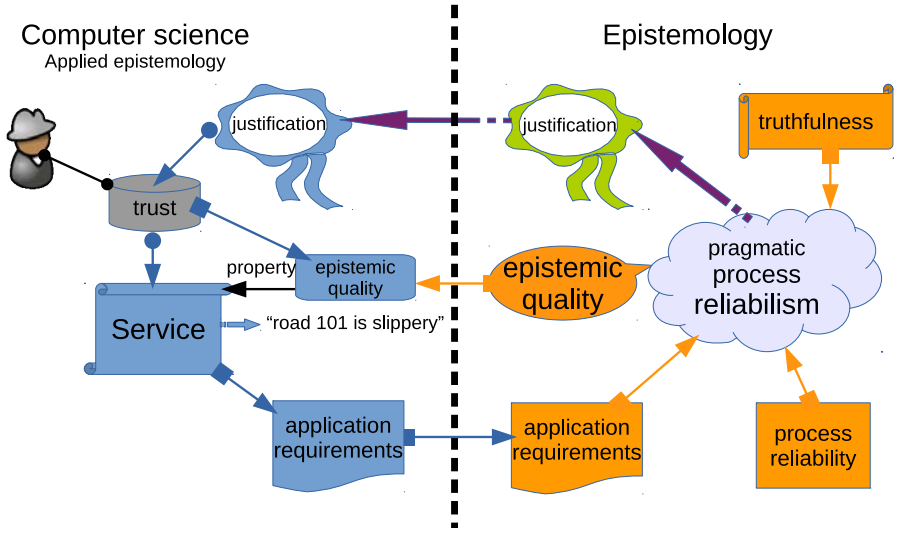

Figure 4.1: Justifiably be trusted.

whether a belief is justified or knowledge is commonly based on quantitative factors (in our case based on the PPR theory).

The concept of "*justifiably be trusted*" begs a question: What kind of justification for trust is plausible? We have already discussed this issue above in Sections 3.5 and 3.7, and now we will establish a connection between *justifiably be trusted* and *the epistemic quality of belief*.

In Section 3.7 we defined trust as follows: *A trusts B to have a property X, and B is trustworthy if it actually has property X, otherwise it betrays A, if A has established a strong confidence based on proper factors that B has property X.* We explicate the above as follows (Figure 4.1):

- The property X of a dependable system B is that the service of the system B either provides users with beliefs that fulfil the epistemic quality (knowledge, justified belief, or belief) expected by users or executes an action that is based on beliefs that fulfil the epistemic quality expected by users and fulfil users' demands.

- A proper factor is the reliability^p of either the process that produces belief or the process that certificates belief.

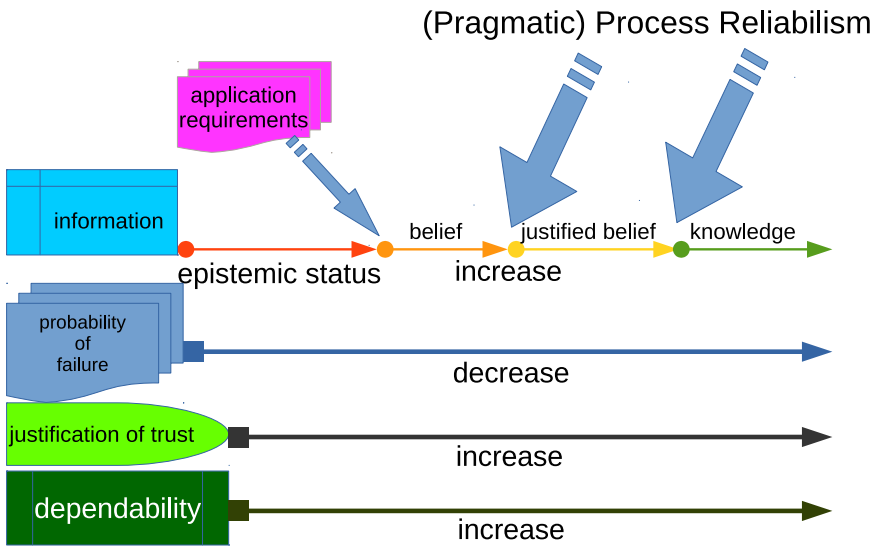

Figure 4.2: High level scheme of context of epistemic evaluation.

- The reliability^p of the process (as the probability of success) can be derived and/or determined from various constituents:

  1. Hardware specifications
  2. Software specifications
  3. Specifications of specific methods and algorithms
  4. Reliability^p measurements
  5. The history of the process (if it is available) producing a service
  6. Certification services
  7. etc.

A high level scheme of the context of the epistemic evaluation is illustrated in Figure 4.2. We claim that a system is justifiably trustworthy only if it carries out actions based on only knowledge and/or justified beliefs[1] and in the case of information services the system shall distribute information with

[1]Of course, there are malicious systems, which are designed to distribute false information and avoid technical service failures that are more frequent and more severe than is acceptable; thus, they are dependable in this sense.

associated metadata expressing the epistemic quality of beliefs (knowledge, justified beliefs or beliefs without any justification). Operating on uncertain (likely false) beliefs seldom produces trustworthy service.[2] We also claim that there is a causal relationship between the epistemic quality of beliefs provided by the system and the service of the system to be justifiably trusted. We assume that in general, the higher the epistemic quality of the beliefs of the system is, the more justifiably trusted the service of the system is.

4.2 Evaluation of Epistemic Quality of Belief

As already discussed in Chapter 3, the epistemic quality of belief depends on the reliability^p of the belief–forming processes of the sources (or of the certification services) and expected consequences of the utilization of belief. There are two factors that shall be evaluated in order to define the epistemic quality of belief:

1. The reliability^p of the belief–forming processes and
2. Expected consequences of the utilization of belief.

Next we discuss briefly these two factors, and a more detailed discussion is in Appendix *Discussions about Reliability of Sources of Beliefs.*

4.2.1 Sources of Beliefs

There are several different kinds of information sources, and we discuss some of their specific features concerning the epistemic quality of belief. We can split ISA$_{bdi}$'s belief–forming processes into two distinct type of processes. First, processes that are external to ISA$_{bdi}$, and second, ISA$_{bdi}$'s internal processes. In this section we assume that ISA$_{bdi}$'s own belief–forming processes are reliable^p enough, and we concentrate on the external processes. We use the scenario of TIS (see page 14) and the belief *"Road 101 is slippery."*. Figure 4.3 illustrates the possible main sources of information, from where ISA$_{bdi}$–A may perceive information either through data communication services or directly using different input mechanism, such as internal application interface, keyboard, mouse, video camera, thermometer, etc. The different cases are summarized in Table 4.1.

[2]The history of Internet has proven this several times.

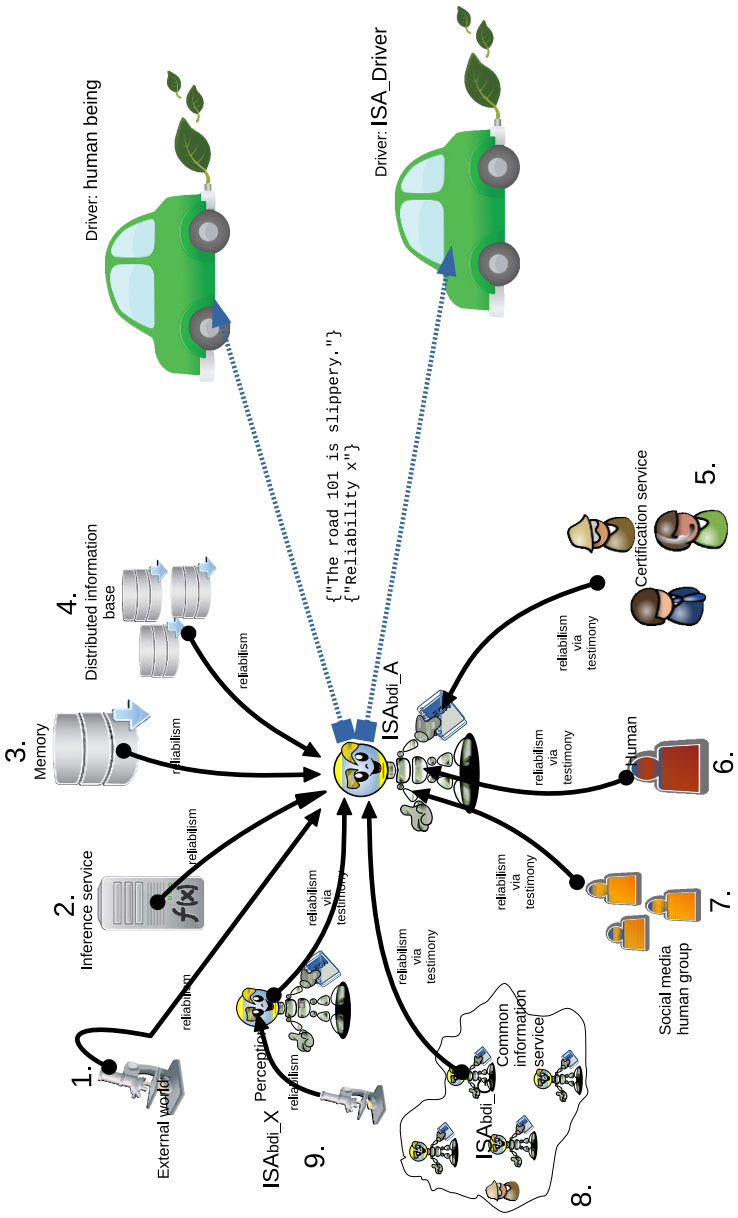

Figure 4.3: Sources of information of ISA.

♯	**Source**	**Theory**[3]	**General Features**
1	Sensor	R	Reliability^p available — it can be obtained from a manufacturer, a certification institute, or user tests.
2	Inference service	R	Reliability^p available — it can be evaluated from the reliability^p of inferring algorithms and their implementations.
3	Memory	R	Reliability^p available — it can be evaluated from the reliability^p of memory specifications and the reliability^p of the belief–forming process. Issues are such as changeable beliefs.
4	Distributed information base	R	Reliability^p available — it can be evaluated from the reliability^p of the specifications of distributed information base. Issues are, for example, public cloud services, changeable beliefs, and in some cases the source of belief.[4]
5	Certification service	R via T	Reliability^p available — certification service evaluate and certifies belief (provides justification); in addition this requires reliability^p of the certification process. Issues are such as the combination of two different reliabilities^p, the certificate of the certification service, and the possibility of autonomous, automatic certification services?
6	Human Being	R via T	Reliability^p is not in general available — it requires using of a certification service or a priori warrant of the human being. Issues are such as difficulties to evaluate reliability^p of a human being and does a warrant of a professional status guarantee the reliability^p of the process.

[3] Adopted epistemic theory: R = reliabilism and T = testimony

[4] We assume that information stored in the distributed information base has been evaluated to be reliable^p.

7	Social media	T	Reliability^p is not available — it requires using of a warrant/certification service or new social media applications, which can manage the demanded metadata and their creation. Issues are such as incompetent human beings, fake persons, crackers, malicious sources of beliefs, etc.
8	Common information service	R via T	Reliability^p is either available (by a priori warrant: warranted service provider) or not available depending on the status of the service. Issues are such as difficulties to evaluate reliability^p of common information service: who does it?
9	Another ISA$_{bdi}$	R via T	Reliability^p either available or not available depending on the ISA$_{bdi}$-X. May require use of a certification service or a warranted ISA$_{bdi}$. Issues are such as difficulties to evaluate reliability^p of a ISA$_{bdi}$ and does a warrant of a '*professional*' status guarantee the reliability^p of the process.

Table 4.1: Summary of sources of belief.

4.2.2 Evaluation of Consequences

The evaluation of the possible consequences of an action carried out by ISA$_{bdi}$ is a difficult task, which comprises problems—such as the presentation of the actual world, specifying relevant counterfactual worlds, the frame problem, and specifying the reliability^p requirements of each counter factual world not to happen—that may not fully be solvable with current computing power and the state of the art in logics and AI. But, we are of the opinion that a lot can be done even today. The evaluation of the consequences is a fully application dependent issue, and as such, it is outside the scope of this thesis. But we give a more detailed example in Appendix *Discussions about Reliability of Sources of Beliefs* how it could be carried out in the case of the belief *"Road 101 is slippery."* (illustrated in Figure 4.4 and in Table 4.2).[5] The key idea is as follows: first, look up relevant con-

[5]The reliability^p numbers in the table are made–up and they are only examples of possible values.

#	Action	Reality - road is	Status	Result	ADC R^p	HDC R^p
1	Declare traffic notice	slippery	possibly correct	success	0.65	0.60
2	Declare traffic notice	not slippery	incorrect	failure	0.35	0.40
3	Do not declare traffic notice	slippery	incorrect	failure	**0.30**	**0.35**
4	Do not declare traffic notice	not slippery	possibly correct	success	**0.70**	**0.65**
5	Declare traffic warning	slippery	possibly correct	success	0.80	0.74
6	Declare traffic warning	not slippery	incorrect	failure	0.20	0.26
7	Do not declare traffic warning	slippery	incorrect	failure	**0.15**	**0.23**
8	Do not declare traffic warning	not slippery	possibly correct	success	**0.85**	**0.77**
9	Declare traffic alert	slippery	possibly correct	success	0.95	0.90
10	Declare traffic alert	not slippery	incorrect	failure	0.05	0.10
11	Do not declare traffic alert	slippery	incorrect	failure	**0.01**	**0.05**
12	Do not declare traffic alert	not slippery	possibly correct	success	**0.99**	**0.95**

Table 4.2: Relevant possible worlds of Traffic Information Service and evaluated reliability^p requirements for declarations.

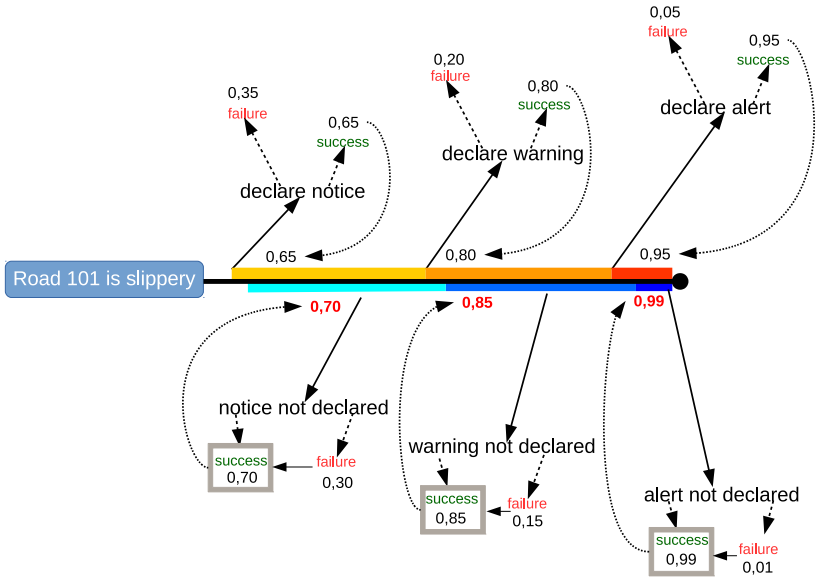

Figure 4.4: Evaluation of consequences.

sequences; second, specify their severity (acceptance of happening); third, specify their reliability^p requirements, and fourth, specify the reliability^p limits for belief, justified belief, and knowledge. A belief may have several epistemic qualities depending on the environments and circumstances, where it is utilized.

In this example there are 12 relevant possible consequences (1 real world and 11 counterfactual worlds) as listed in Table 4.2. For each relevant world a reliability^p requirement is specified based on estimations of users' acceptance of the occurring of consequences. For example, not declaring a traffic alert when the road is slippery (♯11) is not the accepted consequence because it may cause traffic accidents. In this example it is accepted to happen only once (ADC) or five times (HDC) in 100 cases, when the road is slippery. In this example this specifies directly its reliability^p requirement for knowledge.

4.3 Summary of Belief as Dependability Factor

We claimed that a system can be justifiably trustworthy and therefore dependable only if it carries out actions based on only knowledge or justified

beliefs, and in the case of information services the system distributes only knowledge and justified beliefs with associated metadata expressing the epistemic quality of beliefs. We established the connection between *justifiably be trusted* and *the epistemic quality of belief* so that epistemological theories can be used to establish one ground for the evaluation of the dependability of ISA_{bdi} and IDS.

There are several different kinds of information sources, and each one of them has specific features concerning the epistemic quality of belief. A preliminary analysis showed that in many cases the reliability^p of belief–forming processes can be obtained, but there are sources of which reliability^p of belief–forming process is not available. Therefore, there is a requirement for third party warrant/certification services.

The evaluation of the possible consequences of an action—determining relevant possible worlds—is a difficult task. It comprises problems that may not be solvable with current state of art in logics and AI [6], but we claimed that a lot can be done even today. We showed using an example how the evaluation of possible consequences and requirements for the reliability^p of belief–forming processes could be carried out. For a more detailed discussion in the Appendix *Discussions about Reliability of Sources of Beliefs.*

Chapter 5

Enhancement to Dependability Taxonomy

The role of computing systems in the society is rapidly changing towards autonomous intelligent software agents, which are operating more and more in the environment of uncertain information. Therefore, as already mentioned in Section 2.1.2, we argue that some assumptions presented in Section 2.1.2 will not hold in the future. The concepts of system dependability need enhancements in order to be applicable in the environment of the future distributed computing systems based on AI, ISA_{bdi}s and robots.

5.1 Issues of Dependability Taxonomy

Next we point out issues that we need to evaluate more thoroughly. When we deal with intelligent—i.e. autonomous, capable of learning—software agents, which are co–operating with human beings and other ISAs, the following terms are not clear any more:

- Functional specification
- Correct service
- Service failure
- Service outage
- Degraded mode.

The role of the functional specifications of a system is to describe as accurately as possible what the system to be developed should do, so that the designers and programmers of the system are able to implement the

system correctly. The functional specifications are also used in the production phase to reveal service failures. When we deal with a system, which is capable of learning new services (and also to forget old services), it is not possible to define the functional specifications of the system accurately in the way that is required by Laprie's dependability model. It is difficult to predict what kind of new functionalities the system might learn during its production phase. Let us have an example of an avatar, whose original functional specifications state that the avatar's function is to provide legal advices in the domain of e–commerce. However, the avatar happens to learn, when analysing legal issues, also the trustworthiness of various e–commerce companies, and therefore, the avatar begins to provide its customers with the ratings of the trustworthiness of e–commerce companies. Is this a correct service or not? The answer depends on the aims of the avatar. If the aim of the avatar is to offer only legal advices, then the service is not a correct one. But if the aim of the avatar is to learn and serve also new services, then the service is the correct one.

The problems with *correct service*, *service failure*, and *service outage* can be elaborated with the following example: In a nursing institute NI the system function of a nursing robot NR is to transport paralysed patients to a fitness hall for daily physical exercise. However, after having transported a patient P during several months, NR learns that P does not actually want to have physical exercise on Sundays, but wants to be transported to a nearby seaside terrace to watch sailing yachts. Therefore, this Sunday NR does not transport P to the fitness hall, but to the terrace. Now, did NR deliver the correct service or was there a service failure and outage? This depends on the viewpoint: from the viewpoint of NI it was a service failure and outage as it deviated from the patient's rehabilitation plan. But from P's point of view it was the correct service, as it followed P's understanding about what was good for her/him. Therefore, the roles of the users[1] need to be taken into account in other terms than *interaction faults resulting from human errors*.

The *degraded mode* of a system is an interesting concept in the contexts of ISA_{bdi} and DIDS, particularly when ISA_{bdi} deals with different qualities of information, such as belief, justified belief, and knowledge. For example, if ISA_{bdi} has only beliefs—not justified belief or knowledge—to infer a next action to carry out a service to its users, is it the case of the degraded mode of the system? Again, this depends on the aim of ISA_{bdi}. For example, if the aim of ISA_{bdi} is to provide services in all possible circumstances, then

[1]There are two classes of users of the nursing robot: the nursing institute and the patients, and these have different viewpoints on the service.

we can say that ISA_{bdi} operates in the degraded mode. But if the aim of ISA_{bdi} is operate only on knowledge, then we can say that ISA_{bdi} is not in the degraded mode but in an error state.

Laprie's model of dependability assumes that critical computer systems are developed and used by organizations rather than individuals, and there is an expected way of working. Therefore, it is possible to recognize deviations from a correct service and associated system failures [32]. As several examples[2] in the literature of AI and robotics indicate this is no longer the case. The dependability of systems extends beyond the hardware and software into the social and lived experience of the group of various users. Intelligent systems become a part of the self–concept for users, and therefore, it is essential that dependability does not just mean that a system behaves according to the expectations of its designers, but users may have requirements of dependability that are beyond the expectations of the designers of the system [32].[3]

5.2 Attributes

As already mentioned in Section 2.1.2 (see Figure 2.4 on page 19) Laprie's model comprises the following attributes of dependability [12]:

1. *Availability*, which is defined as the readiness for the correct service.
2. *Reliability*^{c}, which is defined as the continuity of the correct service.
3. *Safety*^{c}, which is defined as the absence of catastrophic consequences on the users and the environment.
4. *Confidentiality*, which is defined as the absence of unauthorized disclosure of information.
5. *Integrity*, which is defined as absence of improper system alterations.
6. *Maintainability*, which is defined as the ability to undergo modifications and repairs.

But, these attributes are not sufficient enough for the evaluation of some dependability issues in the environment of future intelligent systems. For instance, in the nursing robot example above, how do we evaluate the system that has learned the new service (transporting the paralysed patient to the seaside terrace to watch sail yachts)?

[2] Autonomous cars, social robots, etc.

[3] For example, this can be the case when users begin to use an intelligent system either in a way or in an environment which designers have not foreseen or expected at all.

- Availability: The system is ready for usage, which was originally assigned to it to be the correct one. But it has inferred based on learning new justified beliefs that the new usage would also be a correct and better one. Therefore, availability is not affected by implementing and executing the new service.

- Reliability^c: The correct service is not interrupted in any way (of course, this depends on the viewpoints, which might result in conflicting opinions.[4])

- Safety^c: The risk of catastrophic consequences has not increased in any significant way.

- Confidentiality: The risk of unauthorized disclosure of information has not increased.

- Integrity: The risk of improper system alterations has not increased in any significant way.

- Maintainability: Maintainability is not reduced in any significant way.

As we can see these attributes do not provide any proper ones to evaluate the dependability of systems that can learn—adapt their behaviour and services—in a changing environment. Hence, we propose three new attributes:

- **Skilfulness**, which we define as the ability to be cognitively skilful to improve an existing service or to develop and to implement a new correct service.

- **Truthfulness**, which we define as the ability to operate on and produce belief that satisfies the epistemic requirements of a correct service.[5]

- **Serveability**, which we define as the ability to provide a correct service in the environment of uncertain, conflicting, (or even contradictory) information. This expresses a kind of sensitivity to the epistemic quality of beliefs.

[4]There are at least two factors: Customers and the length of service episodes. There are two customers of this service: the institute and the patient, and there are two distinguished episode lengths: the length of overall treatment consisting of several physical exercises and the length of the single physical exercise.

[5]We discuss epistemic requirements in Chapter 3.

These three attributes enable us to take into account the social and lived experience of the group of users. Serveability enables us to consider the effects of the epistemic quality of beliefs on the dependability of ISA_{bdi} and DIDS. For example, we can evaluate NR's skilfulness based on the consequences of the service for the patient's well–being. If the patient's well–being is increased when having the new service, then NR's skilfulness is good, otherwise NR needs to be improved in this domain. If NR has based the development and implementation of the new service on knowledge and not on mere belief, then the truthfulness of the nursing robot NR is high. If the new service is considered to be a correct one, then NR's serveability is good, otherwise it is poor.

5.3 Faults

Laprie's model specifies three different groups of faults (Figure 2.4 on page 19) [12]: development faults, interaction faults, and physical faults. A more detailed taxonomy of faults comprises the following list [12]:

1. Phase of creation or occurrence: *development faults* and *operational faults*
2. System boundaries: *internal faults* and *external faults*
3. Phenomenological cause: *natural faults* and *human-made faults*
4. Dimension: *hardware faults* and *software faults*
5. Objective: *malicious faults* and *non–malicious faults*
6. Intent: *deliberative faults* and *non–deliberative faults*
7. Capacity: *accidental faults* and *incompetence faults*
8. Persistence: *permanent faults* and *transient faults*.

We argue that this categorization needs to be enhanced in order to better describe possible faults, and thus to better enable implementing the means of dependability in systems that learn new functions and services. For instance, in the example of the nursing robot NR above, let us assume that the case is a failure.[6] So NR has learned a new behaviour that is considered to be a service failure. In other words, NR has learned a belief that it should not have taken into account when planning and executing actions. How do we categorize this fault?

[6]The viewpoint of the institute.

Fault Class	Valid	Fault Class	Valid
development fault	no	operational fault	yes
internal fault	yes	external fault	yes
natural fault	no	human–made fault	yes–no
hardware fault	no	software fault	yes–no
malicious fault	yes-no	non–malicious fault	yes-no
deliberate fault	no	non–deliberate fault	yes
accidental fault	yes	incompetence fault	yes
permanent fault	yes	transient fault	yes

Table 5.1: Fault classes.

One approach is presented in Table 5.1. The main issues about this approach are as follows: We cannot claim that the failure is the result of a development fault because if we claimed so, it comprised a requirement that either in the development phase all the things that the system is not allowed to learn should be specified, or the system should learn what not to learn. We claim that this is not feasible. It is an operational fault (as it occurred during service delivery of the use phase), but the question is in which phase the fault occurred. Was it in the learning or execution phase? There is no simple answer. This does not help much when designing and implementing the means of dependability. The fault can be either an internal fault (originate inside the system boundary) or an external fault (originate outside the system boundary). Once again this is not very helpful. The fault can also be either a malicious fault (introduced by a human with the malicious objective of causing harm to the system) or a non–malicious fault (introduced without a malicious objective) depending on patient P's psychological objectives. Again, this does not help much when designing and implementing the means of dependability. The same applies to a deliberate fault (result of a harmful thinking) and a non–deliberate fault (introduced without awareness). The fault is not an accidental fault because it is the result of a deliberate reasoning. We can say that it is an incompetence fault. But we argue that this fault class is too undetailed to be properly helpful. It does not specify the kind of incompetence, which would be required when designing and implementing the means of dependability. The fault can be either permanent or transient depending on future learning.

As we can see there is a need for fault classes that can help to manage faults in the context of training, learning and epistemic concepts. In order to properly support the design and implementation of the means of

dependability we need the following fault classes:

- **Training fault**: A trainer or a user has taught a system an erroneous behaviour or a false belief.
- **Learning fault**: A system has learned an incorrect behaviour or a false belief despite proper training or other proper perception.

5.4 Failures

A dependability failure occurs when a system suffers service failures more frequently or more severely than acceptable [12]. We should consider the extent to which a system possesses the attributes of dependability in a relative, probabilistic sense, thus not in an absolute, deterministic sense [11].

There are four viewpoints to characterize a service failure [12]:

1. Failure domain: *content failures, timing failures*, and *erratic failures*
2. Detectability of failures: *signalled failures* and *unsignalled failures*
3. Consistency of failures: *consistent failures* and *inconsistent failures*
4. Consequences of failures on the environment: *minor failures, ... , catastrophic failures.*

We need to extend the above categorization of the service failures to handle better failures happening in social activities. The failure domain does not offer a proper concept for the case such as the one in the above example of the nursing institute. We propose the following enhancement:

- **Action failure**: An action, which a system performs, deviates from the aim of the system.

The detectability of failures does not comprise the failures that users signal to the system for the system to correct itself. We propose the following enhancement:

- **Observed failure**: A failure, which a user of the system observes and informs to the system for the system itself to carry out a corrective action.

The key difference between a signalled (unsignalled) failure and an observed failure is that in the former the system itself is responsible for declaring the failure to a third party for a corrective action and in the latter the user is responsible for declaring the failure to the system for a corrective action.

5.5 Means

When a service has failed because of a training fault a correct fault removal method would be the retraining of the system. In the case of the learning fault there are two options: either to improve the learning algorithm or to change the training of the system. We see that the means of dependability in Laprie's categorization are quite adequate. If we wanted to enhance the means, then we can add the following concepts into the fault–removal (diagnosis, isolation, reconfiguration, and reinitialization):

- **Relearning** in which a system itself replaces a faulty belief or action with a new, correct belief or action.
- **Retraining** in which an external entity trains a system to replace a faulty belief or action with a new, correct belief or action.

5.6 Discussion about New Attributes

These new attributes raise the following question: what is required to achieve them? Skilfulness (being cognitively skilful) requires from the system at least two the following abilities: first, to develop a new kind of methods or to enhance existing methods (action, algorithm, etc.) and second, to be aware of the epistemic quality of information and its meaning to a service to be produced. Our interest is in the second one. In order to have ability of awareness of the epistemic quality of information the system needs the following factors:

- Implementation of the supported epistemic theories
- Application requirements of dependability, that is what are possible consequences and their severity to service users
- Understanding combined probability of failures of various information sources.

Both truthfulness and serveability require from the system the ability to be aware of three factors:

- Epistemic quality of information
- Meaning of different qualities of information to the production of the service

- Possible consequences of the service both to its users and to the environment of the service.

Truthfulness and serveability are directly related to the epistemic quality of the beliefs of the system (perceived, stored, and communicated).

5.7 Problems of Implementing Dependability Concerning Epistemic Quality of Information

Problems to implement dependable distributed systems have already been discussed quite a lot in numerous articles and text books. However, in order to take into account the epistemic quality of information raises, at least, the following new problems:

1. Should there be a directory service that provides data about the sources and their reliability^p of information creation processes (for example, similar to DNS (Domain Name Service) and UDDI (Universal Description Discovery and Integration) but offering reliability^p information)?

2. If so, then how do we implement the domain of the certificated reliability^p of the information creation processes of sources?

3. How do we actually evaluate and warrant/certificate the reliability^p of human beings' information creation processes? Are official licenses (submitted by universities, public administrators, etc.) for executing professions a satisfactory solution, or should there be a real–time evaluation of the results?

4. Are manufacturers and software companies willing to test and inform about the reliability^p of information creation processes of their product and services?

5.8 Summary of Dependability Taxonomy

Jean-Claude Laprie et. al. [82, 83] developed the basic concepts and terminology of dependability in the decades of the eighties and nineties, and they have developed them further in the first decade of the 21st century. However, the expected, future development in the domains of AI, intelligent software agents, and robotics will establish an environment where the established basic concepts and terminology need to be enhanced. We proposed the following ones:

- Three new attributes:
  1. Skilfulness
  2. Truthfulness
  3. Serveability.
- Two new fault classes:
  1. Training fault
  2. Learning fault.
- Two new service failure concepts:
  1. Action failure
  2. Observed failure.
- Two new means:
  1. Relearning
  2. Retraining.

We argue that we are able to better formalize and develop methods to manage and improve dependability of future intelligent distributed systems with these new concepts despite the fact that there are still several problems to be solved. Benefits of the enhanced taxonomy are listed in Table 5.2.

Topic	**Old Taxonomy**	**Enhanced Taxonomy**
Epistemic quality of information	does not take well into account the effects of different epistemic quality of information.	allows better to take into account and understand in design and execution phase the effects of different epistemic quality of information, especially in the cases of service failures and service outages.
Functional specification	does not cover well all aspects of functional specifications, such as in the case of the autonomous learning of new functionalities or services.	enables better understanding the role of functional specification and its limitations, for example, in the context of learning systems.

Topic	Old Taxonomy	Enhanced Taxonomy
Correct service	does not support a correct service to be beyond functional specifications.	enables better understanding the role of functional specifications and takes into account a correct service to be what a system aims at.
Service failure	causes a new, autonomously learned service that is not specified in functional specifications to be a service failure.	allows a new, autonomously learned service not to be classified as a service failure, if the service is according to the aim of a system.
Service outage	does not take well into account different epistemic quality of information that possibility affects the unavailability of an offered service.	allows to take into account the epistemic quality of information in the evaluation of various outages of a required service.
Degraded mode of system	does not take well into account different epistemic quality of information that possibility affects the quality level of an offered service.	allows to take into account the epistemic quality of information in the evaluation of the quality of an offered service. This enables better understanding reasons of the degraded mode of operation in the environment of uncertain information.
Skilfulness	does not specify a needed attribute to manage a situation, where a system learns a new service.	takes into account the evaluation of the capability of a system to improve an existing service or to develop and to implement a new correct service. This enables better understanding the capabilities of a system.

Topic	Old Taxonomy	Enhanced Taxonomy
Truthfulness	does not specify a needed attribute that defines the ability of a system to operate on and to produce beliefs that satisfy epistemic requirements.	specifies the attribute that defines the ability of a system to operate on and to produce beliefs that satisfy epistemic requirements of a correct service. This enables, in turn, to evaluate the dependability of a system in the environment of uncertain information.
Serveability	does not specify a needed attribute that defines the ability of a system to provide a correct service in the environment of uncertain, conflicting information.	specifies the attribute that defines the ability of a system to provide a correct service in the environment of uncertain, possibly conflicting information. This enables, in turn, to evaluate better various levels of the dependability of a system
Training fault	does not specify explicitly a fault class, that would address directly to a fault that is caused by an error in the training of a learning system.	specifies a fault class that enables a better classification of faults in the training phase; this, in turn, improves to carry out right corrective actions.
Learning fault	does not specify explicitly a fault class, that would address directly to a fault that is caused by an error in the learning process.	specifies a fault class that enables a better classification of faults; this, in turn, improves to carry out corrective actions.

Topic	Old Taxonomy	Enhanced Taxonomy
Action failure	does not specify a failure type that describes a failure, which happens in the operation phase when, for example, a system has learned a new service that is the aim of the system and the service faces a failure.	specifies explicitly a failure type that enables a proper analysis in the case when a system faces a failure (does not provide a service that it is aimed at) of an autonomously learned service.
Observed failure	does not specify a failure type which a user informs to a learning system in order to the system to correct the service when the system deviates what is its aim.	specifies a failure type that enables a proper analysis in the case when a learning system faces a failure which is informed to the learning system by its user.
Relearning	does not specify explicitly means which is required to carry out in the case where a system needs to relearn information or operation.	specifies explicitly means which is required to carry out in the case where a system faces a learning fault. This helps, in turn, to carry out right corrective actions.
Retraining	does not specify explicitly a mean which is required to carry out in the case where a system needs to be retrained to obtain a new information or operation.	specifies explicitly the mean which is required to carry out in the case where a system faces a training fault. This helps, in turn, to carry out right corrective actions.

Table 5.2: Summary of improvements on dependability taxonomy.

Chapter 6

Belief Description Framework

In this chapter we introduce Belief Description Framework (BDF) that supports processing of information, beliefs, justified beliefs, and knowledge in order to achieve better dependability of ISA_{bdi} and DIDS. The main objective of this chapter is to demonstrate how epistemic theories of pragmatic process reliabilism can be applied to ISA_{bdi} and DIDS. As John Pollock expressed [106]: *"Implementation achieves two things. First, it requires the theorist to be precise and to think the details through. Philosophers are much too prone to ignore the details, just waving their hands when the going gets rough. That might be all right if the details were mere details and we could be confident that filling them in was a matter of grunt work. But in fact, when philosophical theories fail it is usually because the details cannot be made to work. .... So the first thing implementation achieves is that it requires the theory to be sufficiently precise that it can actually be implemented. It is remarkably common when implementing a theory to discover to your chagrin that there are significant parts of the theory that you simply overlooked and forgot to construct. .... The second thing that implementation achieves is that it provides a test of correctness for theories."*

We use Unified Modeling Language [20, 101] to describe our ideas. A more detailed example of BDF is illustrated in Appendix *UML Diagrams of Belief Description Framework*. In order to demonstrate issues that are application dependent we employ the scenario of TIS.[1]

[1] See page 14. TIS is illustrated in Figures 2.1 and 2.2, and the relevant possible worlds of TIS and the reliability^p requirements for declarations are illustrated in Table 4.2. A more detailed discussion is in Appendix *Discussions on Evaluating Epistemic Quality of Beliefs*.

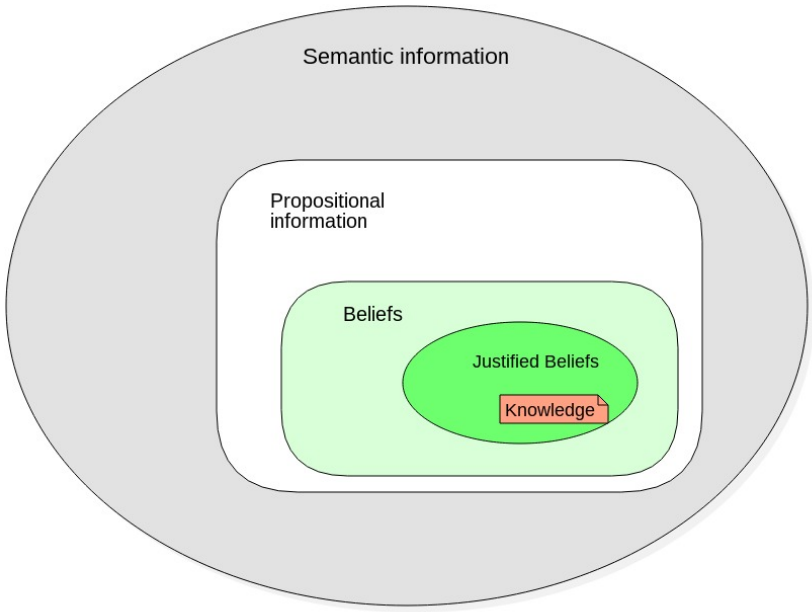

Figure 6.1: Classes of the epistemic quality of information.

6.1 Associations between Epistemic Quality and Software Entities

In general, *semantic information* states something meaningful about something in the context of its usage. Semantic information can be categorized into two classes: non–propositional (such as photographs, paintings, and music) and propositional. Hereinafter, we use the term *information* to refer to propositional information. As discussed in Chapter 3 there are four concepts to describe the epistemic quality of an information stated by a proposition: *information, belief, justified belief,* and *knowledge* (Figure 6.1).

Next we introduce the associations between epistemic theories of these concepts and software entities. *Information* has relationships with several other concepts (Figure 6.2), which affect the determination of the meaning of *information* and the epistemic quality of *information*. The concepts are the following ones:

1. Proposition: Any statement that expresses linguistically a meaningful claim about something and is expressed in a way that an agent can process it properly in a required context.

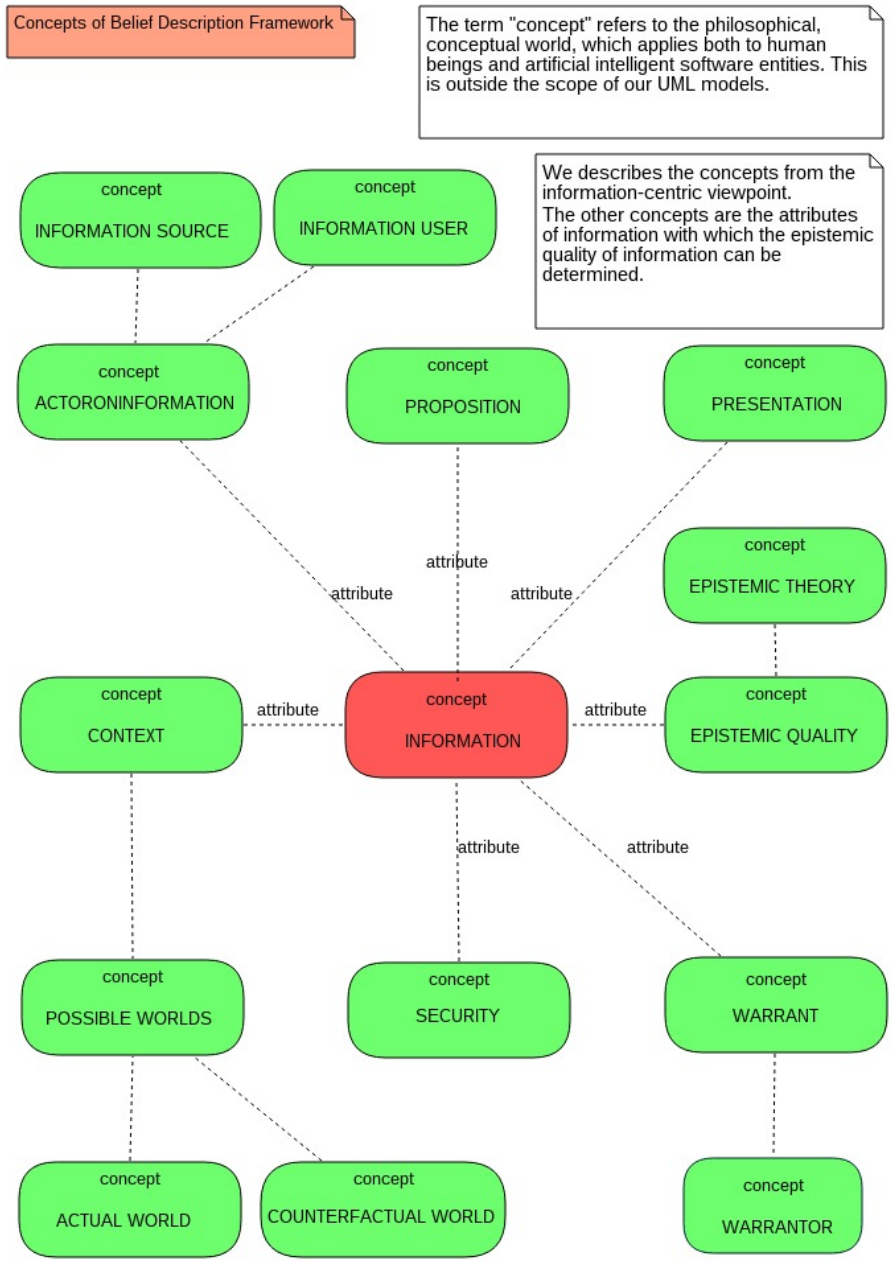

Figure 6.2: Associations of information.

2. Presentation: Any linguistic method to express *information* in the way that *information* users are able to process *information* and the semantic of *information*.

3. Epistemic quality: The property of *information* that expresses the epistemic rating of *information*; that is how well *information* is assumed to correspond to the state of an actual world.

4. Epistemic theory: Philosophical theories of belief, justified belief and knowledge, which are used to specify the epistemic quality of *information*.

5. Warrant: The property of *information* that expresses the assurance of the epistemic quality of *information*.

6. Warrantor: Any epistemic agent that evaluates the epistemic quality of *information* using a reliable^p process and has been itself evaluated by commonly/publicly acceptable methods.

7. Security: The measures that protect *information* against various threats.

8. Context – possible worlds [94]: The environment of *information* that consists of different worlds, where a world is the "limit" of a series of increasingly more inclusive situations. Situations are structured collections of (physical) objects.

   (a) Context – actual world: The environment of *information* that is believed to be a real world.

   (b) Context – counterfactual world: The environments that are believed to be possible situations of *information*, if some nonfactual thing would be valid.

9. Actor on *information*: Any epistemic agent that either produces propositional *information* or uses propositional *information* to deduce actions to be executed.

   (a) Actor on *information* – *information* source: Any epistemic agent that produces propositional *information* that is presented in forms in which *information* users are able to process in a meaningful way.

   (b) Actor on *information* – *information* user: Any epistemic agent that utilizes propositional *information* in order to achieve its aims.

The lower part of Figure 6.3 represents the main information–related components of Virtual Machine Functionality.[2] Now, we can give an expression to *information* (Figure 6.4), and we can instantiate the *information* class as an object (Figures 6.5 and 6.8). We define these concepts to be attributes of *information*. We model the concepts using UML stereotypes (Figure 6.3).

As discussed in Section 3.4, belief is a propositional attitude,

1. which is the state of having an opinion about something to be the case;

2. which is created by its actual and potential causal relations to sensory stimulations, behaviour, and/or other propositional attitudes; and

3. the representation of which—structured if necessary—is stored in a linguistic form.

Based on the above definition an epistemic agent has a belief when the following presumptions are satisfied:

1. The epistemic agent holds a proposition stating the object of the belief.

2. The information object, with which the proposition is associated, is instantiated based on a causal relation to the epistemic agent's perception (perceptions), and/or reasoning from other propositional attitudes.

3. The epistemic agent is cognizant of the identity of an *information* source and/or an *information* warrantor.

4. The epistemic agent is cognizant of the reliability^p of *information* based on either an *information* creation process or an *information* warranting process.

5. The reliability^p of the *information* creation/warranting processes assures *information* likely to be the case.

A belief is the state of an epistemic agent in which there is *information* object (Figure 6.5) stored in the epistemic agent's database of beliefs; thus, the epistemic agent is in the state of having an opinion about something (stated by object *proposition*) likely to be the case.

[2]See Section 3.1. To implement a fully operational Virtual Machine Functionality requires several other components, and it is outside the scope of this thesis.

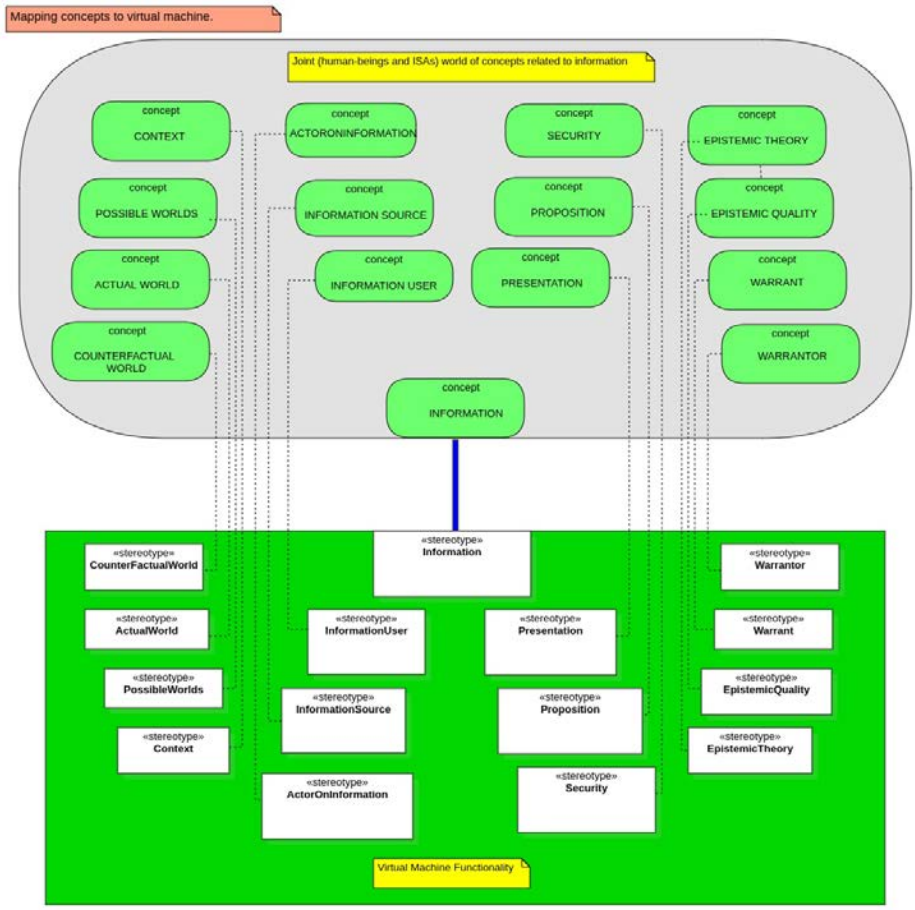

Figure 6.3: Information concepts instantiated as stereotype classes of virtual machine functionality.

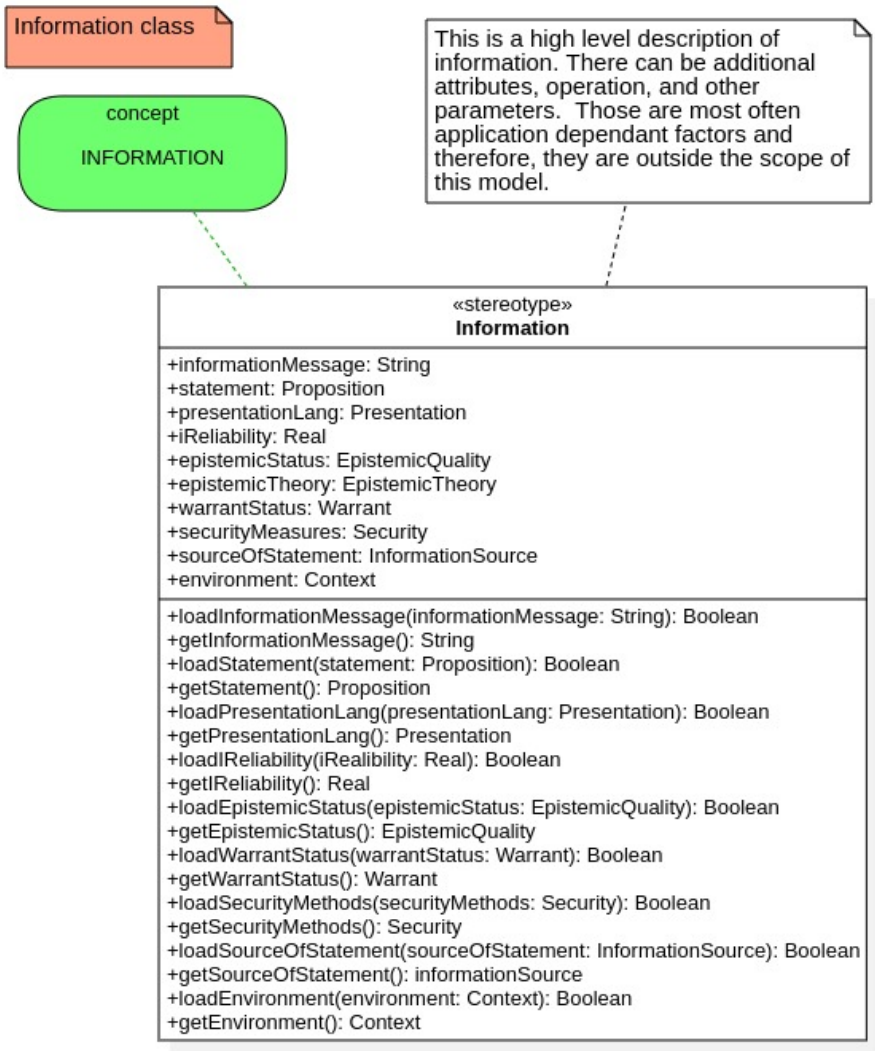

Figure 6.4: Information class.

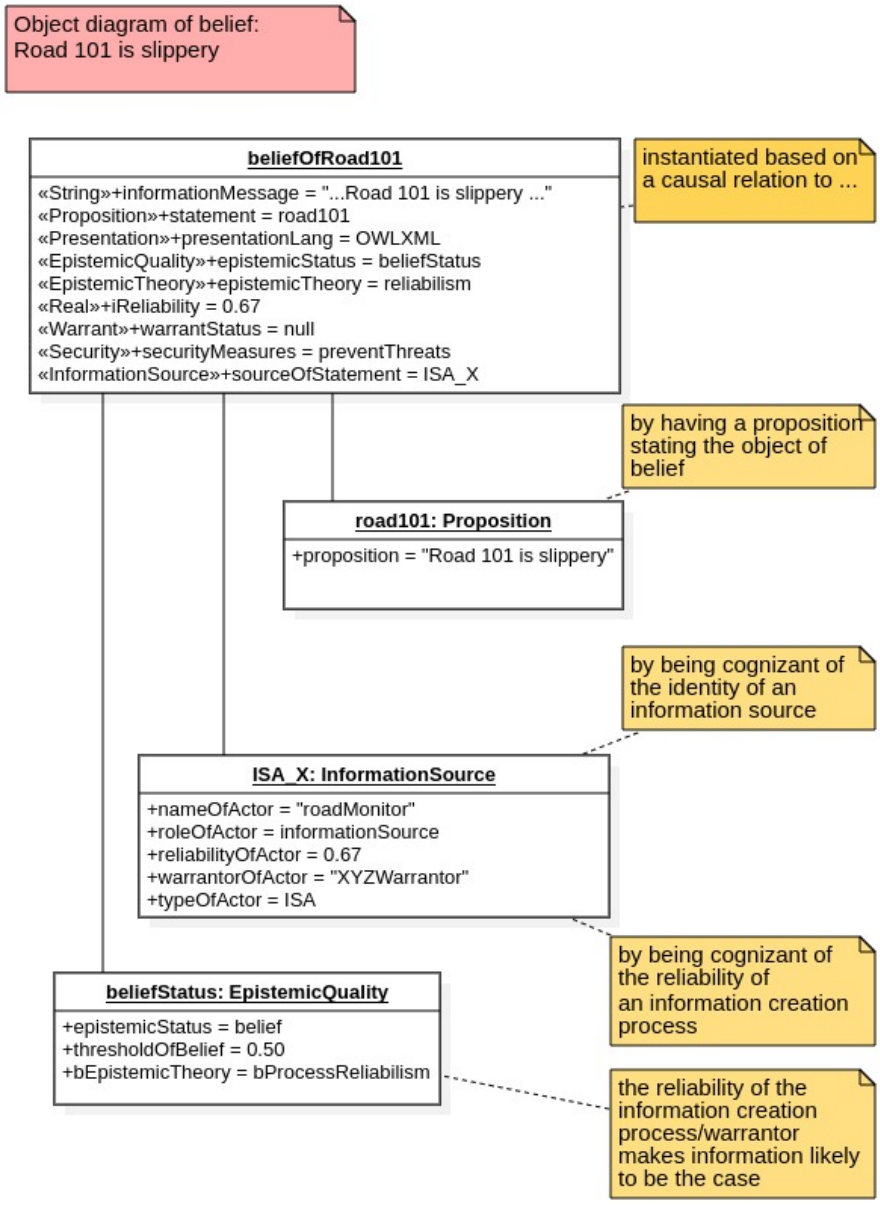

Figure 6.5: An example of a belief object.

The first requirement is satisfied by the object that is an instance of the *Proposition* class and states a proposition. The second requirement is satisfied when the object that is an instance of the information class is instantiated based on a causal relation to an epistemic agent's activity related perceptions or reasoning from other propositional attitudes. The third requirement is satisfied by the object that is an instance of either the *InformationSource* class or the *Warrant* class and the object states the source of *information* and/or *warrant*. The fourth requirement is satisfied by the above object having an attribute *reliabilityOfActor* (and/or *reliabilityOfWarrantor*) that states the reliability^p. The fifth requirement is satisfied by the object that is an instance of the *Belief* class (stereotype EpistemicQuality), and the object states the threshold of *information* (attribute thresholdOfBelief) to be a belief. In general, the threshold shall be greater than 0.50.

We defined justified belief in Section 3.5 as follows: An epistemic agent has justification for hers/his/its belief *that p* if,

1. The agent believes p to be true;
2. The agent's belief *that p* was produced through a reliable^p belief–forming process or warranted by a reliable^p information warrantor; and
3. The reliability^p of the belief–forming process and/or the information warranting process is adequately high for the requirements set by the contextual factors in the environment where the agent utilizes the belief *that p*.

Based on the above definition an epistemic agent has a justification for its belief when the following conditions are satisfied:

1. The agent holds the proposition p;
2. The agent is cognizant of the identity of
   - an information source and/or
   - an information warrantor;
3. The agent is cognizant of the reliability^p of
   - a belief–forming process and/or
   - an information warranting process;

4. The agent cognizant of the requirements of the epistemic quality set by a usage of belief; and

5. The reliability^p exceeds the requirements.

A justified belief is the state of an epistemic agent in which there is *information* object (Figure 6.6), which satisfies the above listed requirements, stored in the epistemic agent's database of beliefs; thus, the epistemic agent is in the state of having an opinion, which has justification, about something likely to be the case.

The first requirement is satisfied by the object that is an instance of the *Proposition* class and states p. Likewise in the case of belief, the second requirement is satisfied by the object that is an instance of either the *InformationSource* class or the *Warrant* class, and the object states a source of *information* and/or *warrant*. The third requirement is satisfied by the above object having an attribute *reliabilityOfActor* (or *reliabilityOfWarrantor*) that states the reliability^p. The fourth requirement is satisfied by the object that is an instance of the *EpistemicQuality* class and states a threshold of *justifiedBelief* for a belief to be justified. The threshold can be reasoned from the objects *possibleWorlds—actualWorld and counterFactualWorlds*. The fifth item in the list is satisfied when the parameter iReliability is higher than the parameter thresholdForJustifiedBelief.

We defined in Section 3.6 that an epistemic agent knows *that* p if and only if

1. p is true;

2. The agent believes p to be true;

3. If the agent were to believe *that* p, p would not be false;

4. The agent's belief *that* p was either produced by a reliable^p belief–forming process or warranted by a reliable^p information warranting process;

5. The reliability^p of the process either exceeds or is equal to the reliability^p requirements of the actions,

   (a) where the agent utilizes the belief p and

   (b) which are set by the expected consequences of the actions.

The implementation of the anti–Gettier safety condition (the 3^{rd} one in the list above) is problematic because the concept of nearby worlds is

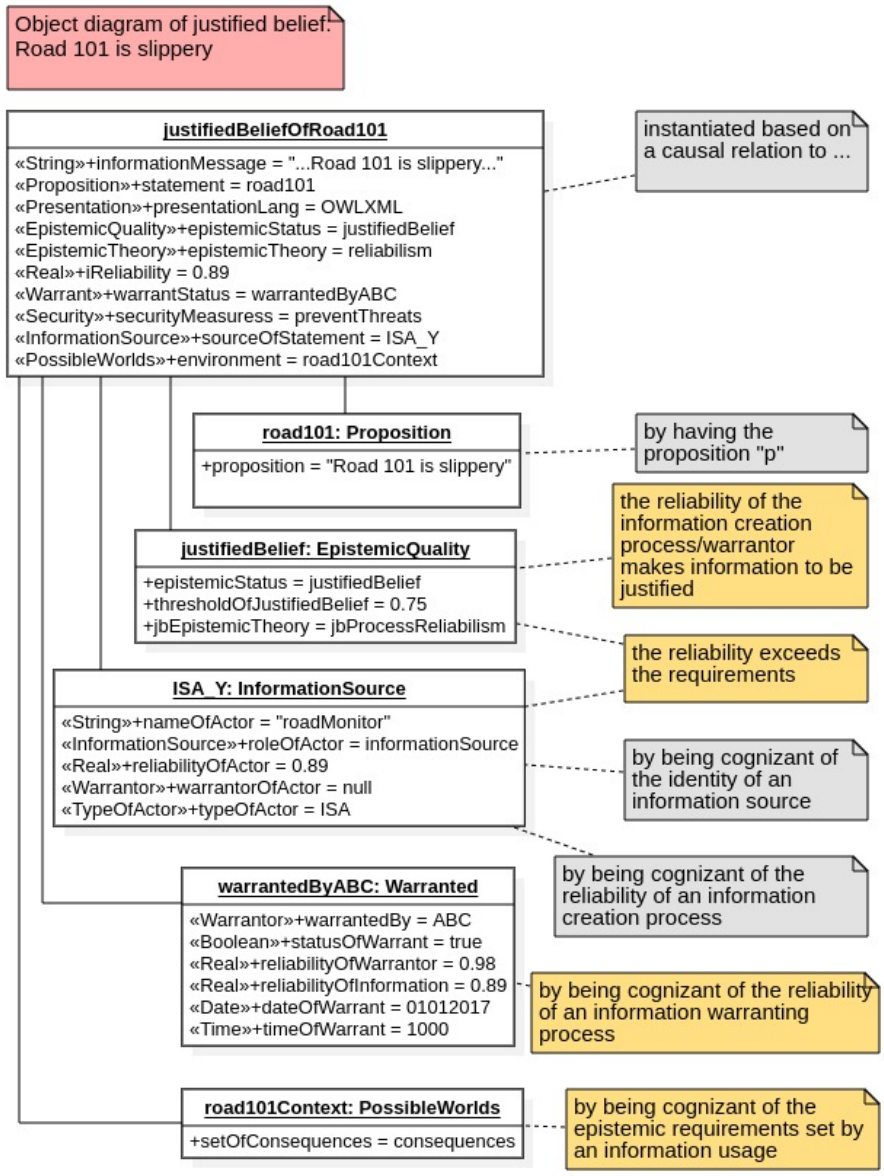

Figure 6.6: An example of a justified belief object.

not sufficiently exact.[3] It could be implemented using the approach of possible worlds and relevance. However, we are of the opinion that the implementation of the safety condition is not necessary for the dependability of ISA_{bdi}. This is because of the following reasons: First, if the reliability^p of a belief–forming process fulfils the requirements of knowledge, then it does not affect the dependability whether the belief is achieved by luck or not. In this sense, luck has no role, here. Second, if p is true, implementing the anti–Gettier condition does not actually increase the dependability.

Based on the above we can argue that an agent knows *that p* when the following conditions are satisfied:

1. p is true;
2. The agent is cognizant of the truth of p;
3. The agent has justification for p;
4. The agent is cognizant of the identity of
   - an information source and/or
   - an information warrantor;
5. The agent is cognizant of the reliability^p of
   - a belief–forming process and/or
   - an information warranting process;
6. The agent is cognizant of the epistemic requirements set by the possible consequences of the usage of information;
7. The reliability^p of the belief–forming and/or warranting processes exceeds the requirements.

Knowledge is the state of an epistemic agent in which there is *information* object (Figure 6.7), which satisfies the above listed requirements, stored in the epistemic agent's database of beliefs; thus, the epistemic agent is in the state of having an opinion, which has required justification for the agent to know, about something to be the case.

The first requirement is problematic because as discussed in Section 3.3 *truth* itself is a difficult concept. However, we argue that the requirement of *"p is true"* is satisfied by a correspondence relation and either the

[3] In addition, the concept of nearby worlds—what are actually relevant nearby worlds—is very much dependent on an application and may vary quite a lot.

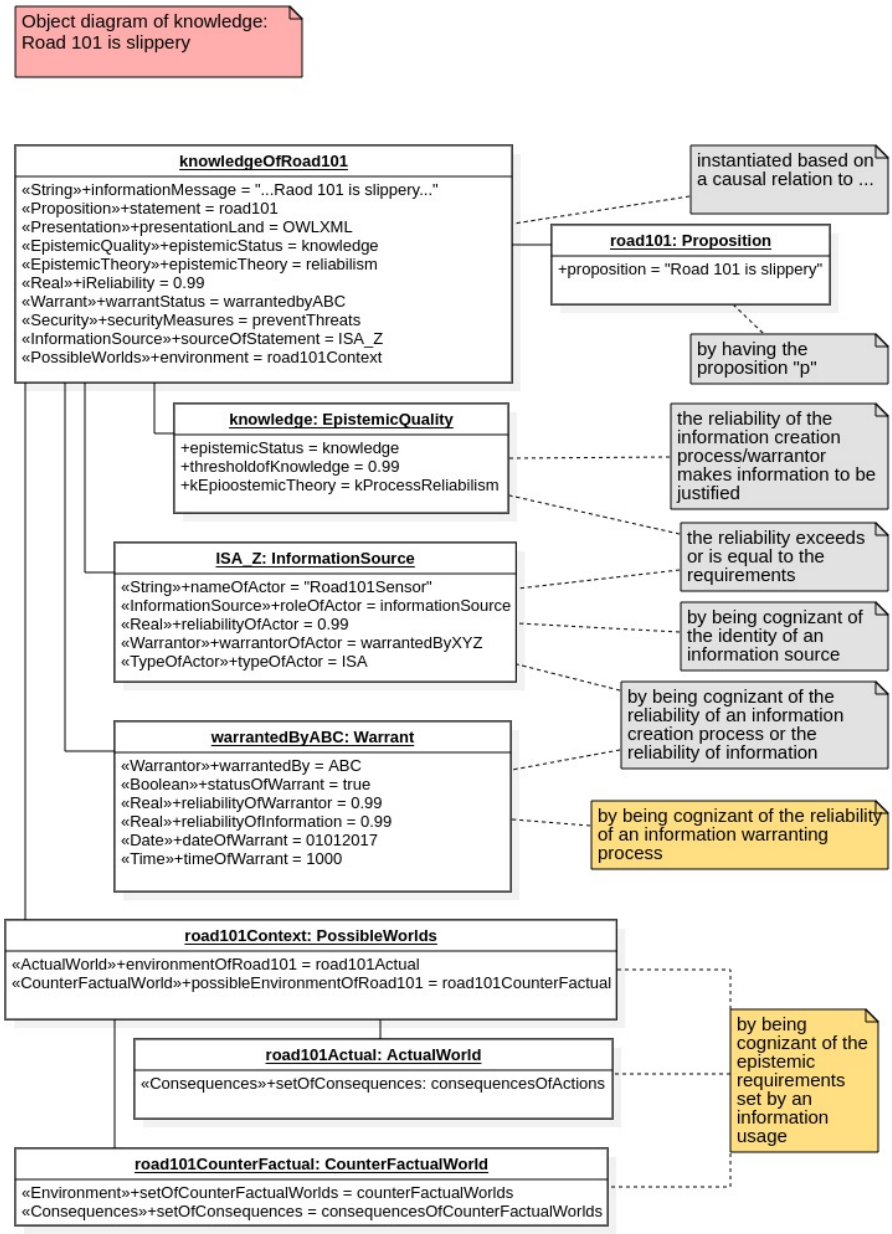

Figure 6.7: An example of a knowledge object.

very high reliability^p of the belief–forming process or a warrant by the very high reliability^p of warranting process (stated by the value of the *thresholdOfKnowledge* attribute)[4], and the agent is cognizant of them. And the reliability^p (parameter iReliability) equals or exceeds the value expressed by the parameter *thresholdOfKnowledge*. The third requirement is already discussed above in the context of justified belief. The fourth requirement is satisfied by the object that is an instance of either the *InformationSource* class or the *Warrant* class, and the object states the source of *information* and/or *warrant*. The fifth requirement is satisfied by the above object having an attribute *reliabilityOfActor* (or *reliabilityOfWarrantor*) that states the reliability^p. The sixth requirement is satisfied by the object *knowledge* that is an instance of the *EpistemicQuality* class, and it states a threshold for a piece of *information* to be knowledge. And the threshold is determined from the objects *possibleWorlds—actualWorld and counterFactualWorlds*. In the case of knowledge the value of the threshold must be significantly higher than in the case of justified belief, and high enough so that there is very little doubt about the proposition to be true. The value of the *thresholdOfKnowledge* attribute must be derived from the analysis of required trustworthiness based on the relevant possible worlds with which the information object is associated. The last item in the list is satisfied when the parameter iReliability is equal or higher than the parameter thresholdForKnowledge. As an example Figure 6.8 illustrates an epistemic agent's knowledge *"Road 101 is slippery"*.

6.2 Requirements for BDF

In this section we introduce the main requirements for BDF at a general level. The objective of the requirements is to increase understanding of BDF and not to specify exact implementation requirements of BDF. We use the scenario of TIS presented in Section 2.1.1 (Figure 6.9)[5] to clarify requirements. The requirements are not bound to any solution technologies. There are five major components:

1. Information,
2. Information source,
3. Information processing including perceiving, evaluation, inferring, and distributing (or acting based on information),

[4]We are of the opinion that fallibility is an option even in knowledge.
[5]See page 14 and Figure 2.3.

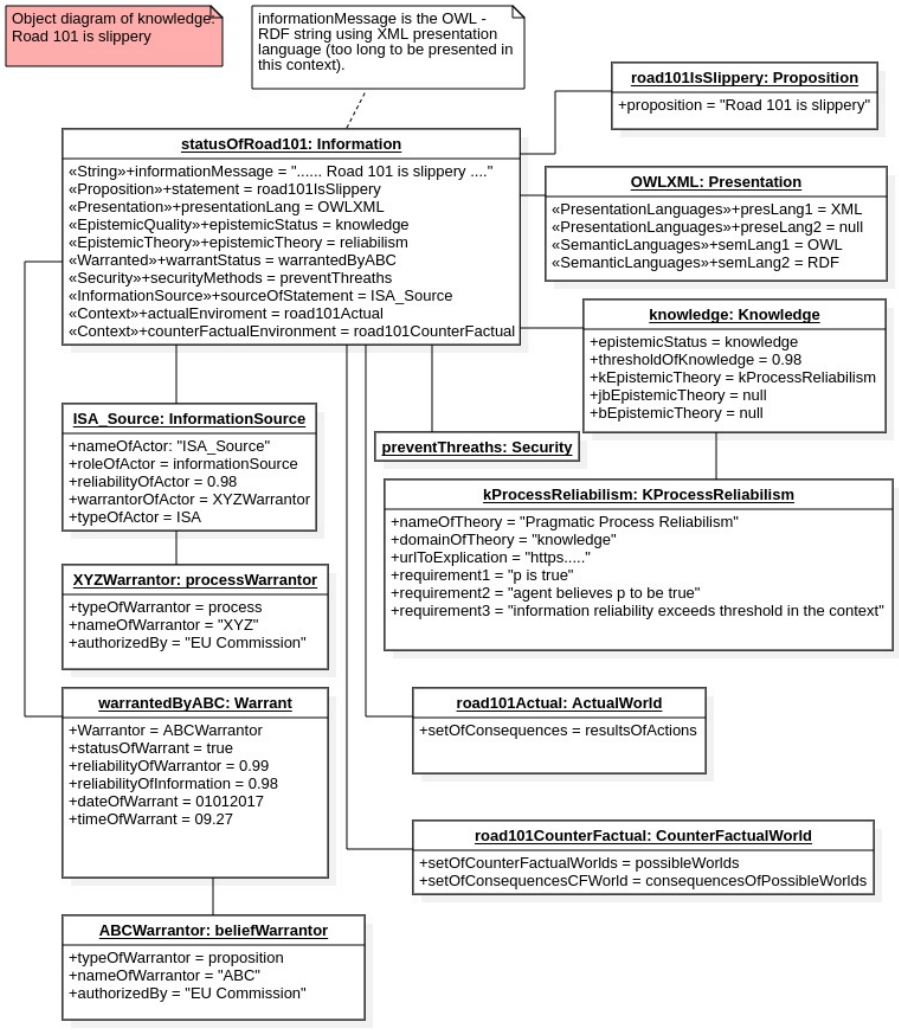

Figure 6.8: Information object structure of knowledge.

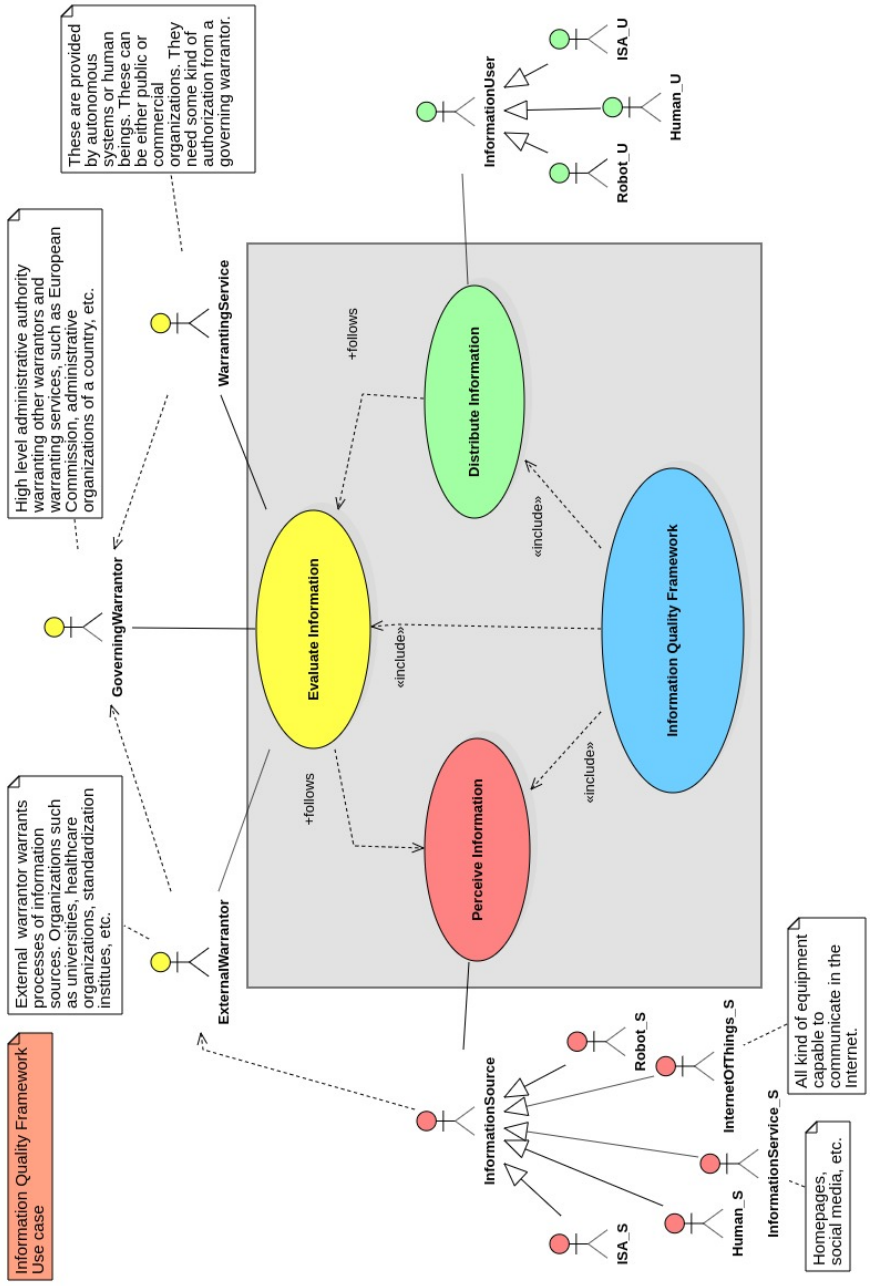

Figure 6.9: An example of use case.

4. Information warrant, and
5. Relevant possible worlds.

Information

As discussed in Section 6.1 information states something meaningful about something in its context. Information never exists in isolation, but it has associations with several other classes. The main requirements of information are the following ones:

InFoReq-1 Information must be semantically meaningful in the contexts of epistemic agents, where information is processed.

InFoReq-2 Information must have a linguistic presentation.

InFoReq-3 The semantics of information shall be embedded using jointly agreed representations of semantics.

InFoReq-4 The representation of semantics shall support the logics that are required to reason about information.

InFoReq-5 Information must be associated with its contexts.

InFoReq-6 Information shall have an epistemic classification.

InFoReq-7 Information must be associated with its source and/or warrantor.

Information Source

As discussed in Section 4.2.1 and illustrated in Figure 6.9 there can be several different kinds of information sources, whose reliability^p of information creation processes varies a lot. Therefore, the source of information affects strongly the evaluation of the epistemic quality of information. The main requirements of an information source are the following ones:

InSoReq-1 The source of information must have a unique identity in her/his/its social context of information.

InSoReq-2 The source of information should have the reliability^p of its information producing process evaluated by an authorized warrantor.

InSoReq-3 The source of information should embed its identity, a reliability^p value, and the identity of warrantor with information messages.

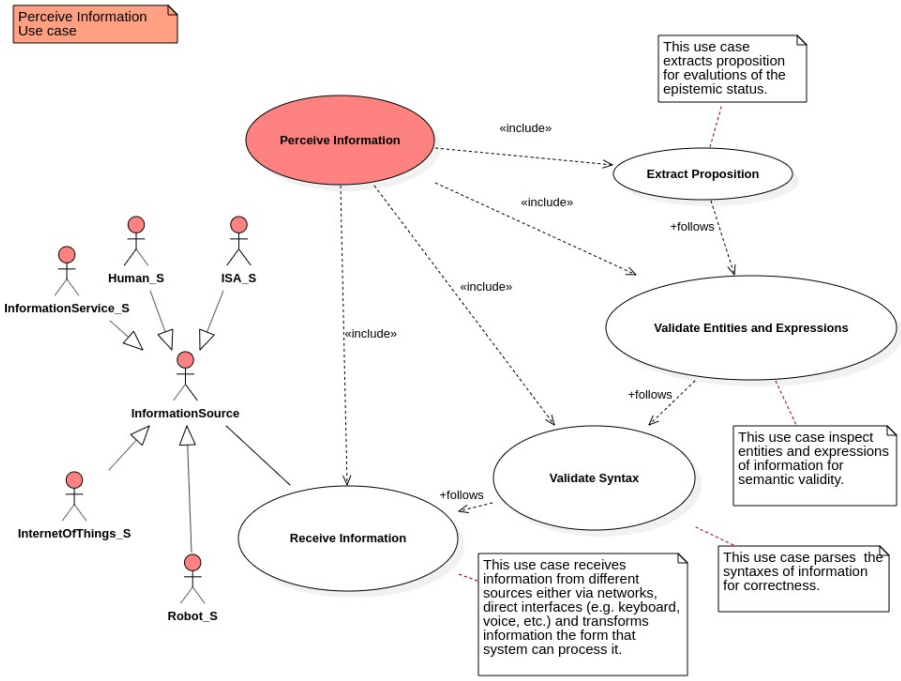

Figure 6.10: Use case: perceive information.

InSoReq-4 The source of information should associate the date and time of the creation of information with information message, if information is contingent.

InSoReq-5 The source of information should be benevolent.

Information Processing

Information processing comprises at least three important domains of activities, which are perceiving (Figure 6.10), evaluating and inferring (Figure 6.11), and carrying out actions (distributing information, selecting the next action to be executed, etc.) as illustrated at a high level in Figure 6.9.

The main requirements of information processing are the following ones:

InPrReq-1 Information processing shall aim at the highest possible epistemic quality of information.

InPrReq-2 Perceiving of information must validate the presentations (syntaxes and expressions) of information.

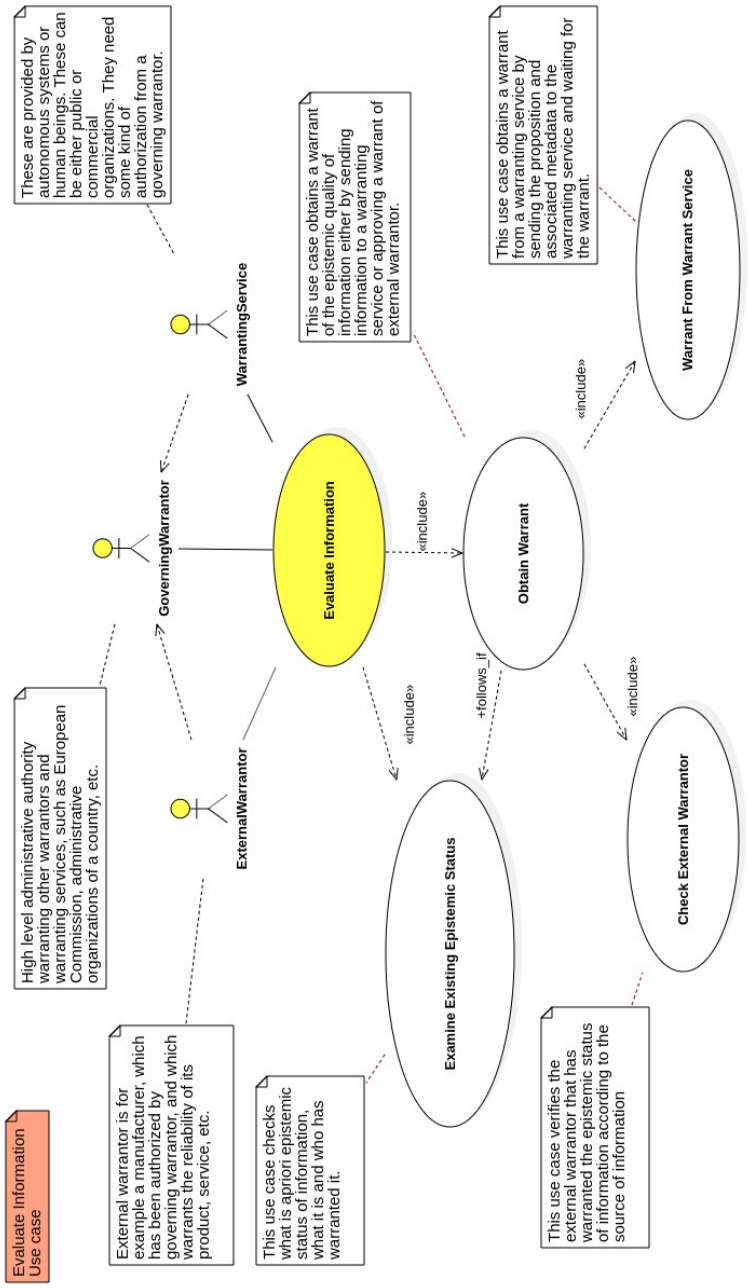

Figure 6.11: Use case: evaluation of information.

InPrReq-3 Information processing must operate only on valid representations and semantics of information.

InPrReq-4 Information processing shall follow the rules of logics supported by the semantic representation language of information.

InPrReq-5 Information processing must maintain the security of information.

InPrReq-6 Information processing shall operate on information in its proper contexts.

InPrReq-7 Information processing shall associate all relevant context factors with information.

InPrReq-8 Information processing shall express all relevant context factors of information.

InPrReq-9 Information processing should be able to manage relevant possible worlds of the results of utilizing information (that means relevant consequences of the actions possibly carried out based on information).

Information Warrant

The main objective of using a warrant service is to obtain justification for information. The warrant service either evaluates the reliability^p of an information creation process or defines the reliability^p of information using other possible factors (e.g. other supporting knowledge, no defeaters, etc.). However, there are instances, where the evaluation cannot be done reliably^p enough because the required information is not available. The warrant service needs to know the reliability^p of its own warrant process.

The main requirements of a warrant service are the following ones:

InWaReq-1 Warrant processing must operate only on valid representations and semantics of information.

InWaReq-2 Warrant processing shall follow the rules of logics supported by the semantic representation language of information.

InWaReq-3 Warrant processing must maintain the security of information.

InWaReq-4 The warrant must express either the reliability^p of an information forming process if available or the reliability^p obtained other valid methods.

InWaReq-5 The warrant must express the base of the reliability^p value; whether it is the reliability^p of the information creation process or the reliability^p computed based on other factors.

InWaReq-6 The warrant shall be expressed as a (probability) value between 0.0 - 1.0, where the value 0.0 indicates total unreliability^p and the value 1.0 full reliability^p.

InWaReq-7 Warrant of information (reliability^p and warrantor) should be available to an epistemic agent that processes information.

InWaReq-8 The warrant must be obtained only from a trustworthy warrantor.

InWaReq-9 The warrantor should announce its services in a directory service.

InWaReq-10 The warrantor should be capable of processing required semantic representations.

InWaReq-11 The warrantor should evaluate or be knowledgeable of its own reliability^p.

InWaReq-12 The warrantor should be authorized by a high level governmental authorizer.

Possible Worlds

The main objective of the possible worlds approach is to model different, relevant and likely outcomes of actions that would possibly be results if the actions were carried out based on an information usage. Possible worlds are used to evaluate the severity of the results and based on the severity to determine the reliability^p requirements. The main requirements of relevant possible worlds are the following ones:

PoWoReq-1 A set of possible worlds must comprise an actual world and relevant counterfactual worlds.

PoWoReq-2 The actual world must be the world the epistemic agent happens to inhabit.

PoWoReq-3 The model of the actual world is the model that the epistemic agent believes to be the surrounding world.

PoWoReq-4 Relevant counterfactual worlds should be considered to be possible worlds that are likely and not–too–distant alternatives of the actual world.

PoWoReq-5 The counterfactual worlds shall state dependencies of whether, when, and how one event occurs on whether, when, and how another event occurs [95].

6.3 Specifications of BDF

In this section we introduce ideas for the specifications of BDF in order to highlight possible solutions to manage the epistemic quality of information. There are several possible ways to implement BDF, and therefore, there can also be several different specifications of BDF. Our specifications of BDF are described using UML in more detail in Appendix *Belief Description Framework*.

6.3.1 Classes and Objects

There are several (stereotype) classes that specify information and its context (Figure 6.12). As already mentioned, we use stereotypes in order to highlight that there are specific epistemic–related concepts which are the basis of the classes. The *Information* class (Figure 6.4 on page 159) is the central point of BDF. The Information class defines the environment—a kind of information ecosystem—of a proposition that states something meaningful about something. It has relationships (dependency, association, generalization, or realization) with other classes: Proposition, Presentation, EpistemicQuality, Warrant, Security, Context, and ActorOnInformation. It also specifies operations of each class. The Proposition class specifies a proposition to be a string. An instance of the Proposition class states a proposition that the information is about. The Presentation class specifies the used representations of *Information* including languages of specifying syntaxes and semantics. An instance of the Presentation class expresses languages that are used to represent information (for example, in information exchange). The EpistemicQuality class specifies the epistemic quality of *Information*. It specifies a proposition to be either info, belief, **justified-belief**, or **knowledge**. It also specifies adopted epistemological theories. An instance of the EpistemicQuality class (Figure 6.13), such as Info, Belief,

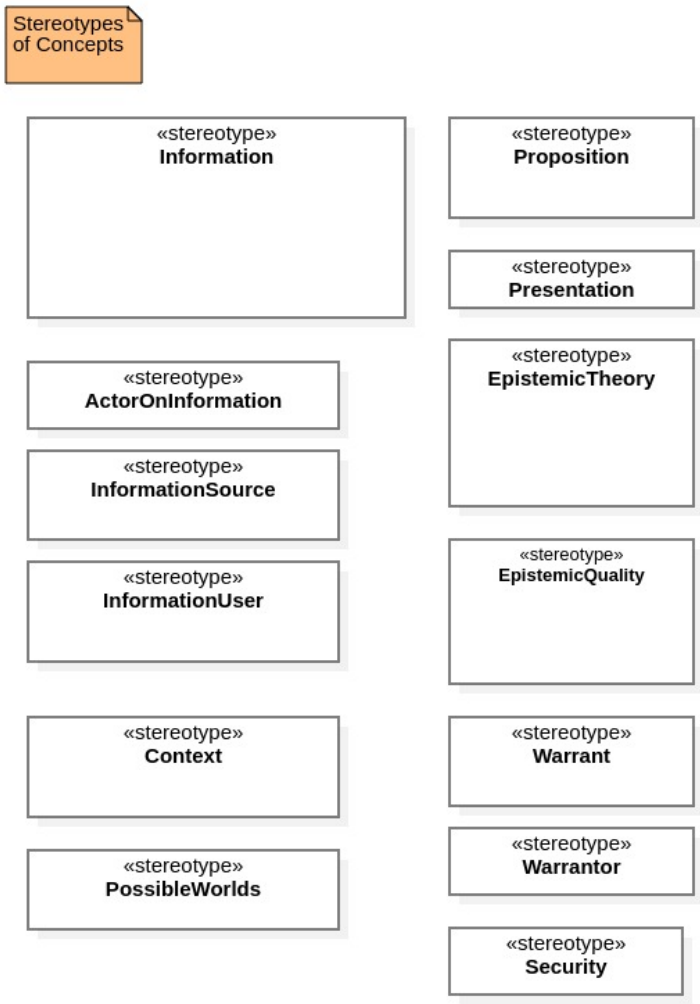

Figure 6.12: BDF classes.

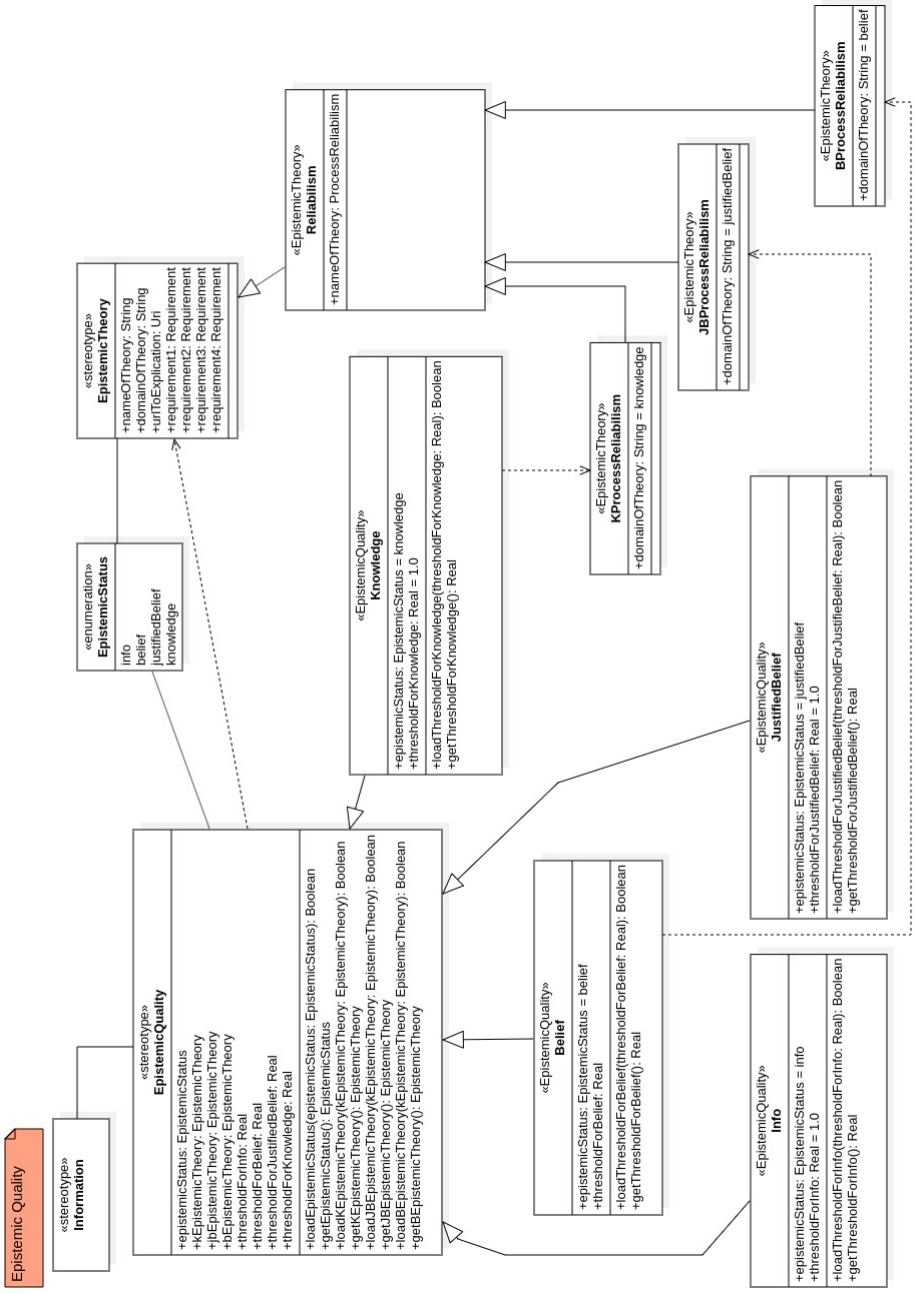

Figure 6.13: BDF epistemic quality class.

JustifiedBelief, and Knowledge, describes the epistemic quality of information. The Warrant class specifies both the status of a warrant and the reliability^p of a warrant. It also specifies the identity and type of a warrantor. An instance of the Warrant class states the situation concerning a possible warrant. The Security class specifies required security measures for *Information*. An instance of the Security class describes the security factors, such as non–reputation, tampering, integrity, deception, and falsification concerning the information.[6] The Context class comprises a subclass PossibleWords that specifies a possible environment of Information including an actual world and counterfactual worlds that are possible outcomes of actions based on Information. This class is instantiated when defining the final epistemic quality of Information by evaluating the severity of each possible world. The ActorOnInformation class specifies the type of agent (TypeOfActor enumeration class) to be human, robot, ISA, informationService, or IOT. It also specifies whether an agent is a source or a user. An instance of the ActorOnInformation class states a source of information or a user of information and some of their features.

As an example of knowledge, Figure 6.14 illustrates one possible object of the Information class that states justified belief "Road 101 is slippery". An instantiation of the JustifiedBelief class specifies that the object in question is justified belief, and it also specifies the threshold of reliability^p for justified belief and an obeyed epistemic theory. If *Information* is knowledge, then the object is associated with a Knowledge object, and if Information is mere belief, then the object is associated with a Belief object.

6.3.2 Collaboration

In general, BDF is a component that is utilized in several phases of the information processing. In our scenario of TIS there are at a high level three phases that are the following ones: perceive, evaluate, and utilize (distribute) information.

ActorOnInformation is the main thread that controls all the activities in the *information* processing. It creates, activates, and destroys required activity–specific threads, such as InformationReceiver, SyntaxValidator, ExpressionValidator, EvaluateInformation, ExamineAprioriEpistemicStatus, ObtainWarrant, ContextEvaluator, DistributeRequisite, and InformationDistributor.

The PerceiveInformation phase consists of the following agents and actions: InformationReceiver obtains a message from a source, extracts *in-*

[6] These are outside the scope of this thesis.

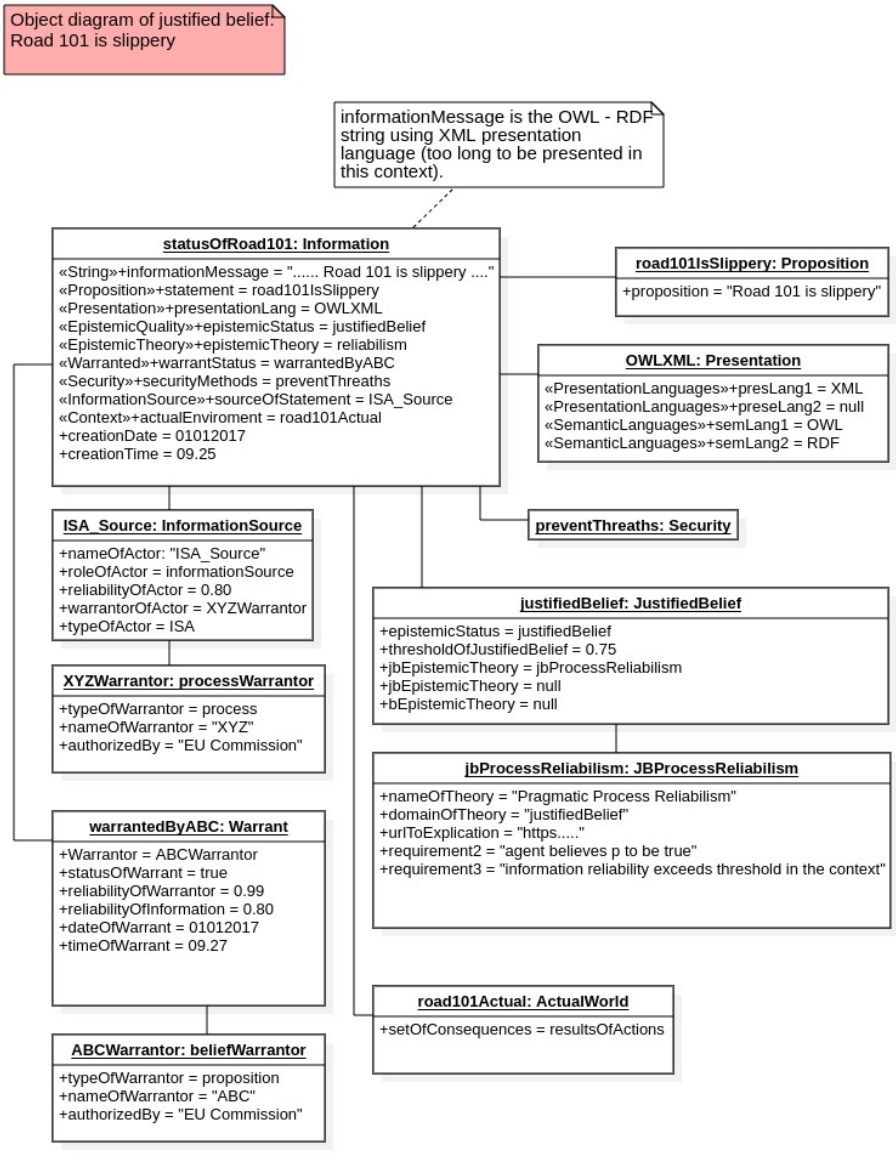

Figure 6.14: BDF instance of information class: justified belief.

formation from the message. SyntaxValidator validates the syntaxes and ExpressionValidator validates the expressions in which *information* is represented.[7]

The EvaluateInformation phase consists of the following agents and actions (Figures 6.15, 6.16, and 6.17): ExamineAprioriEpistemicStatus, ObtainWarrant that comprises SearchWarrantor, and Warrantor. ExamineAprioriEpistemicStatus evaluates the a priori epistemic quality and determines whether a warrant is needed or not. The key idea of evaluating the a priori epistemic quality is to have a first level understanding of the epistemic quality of *information*, which affects the further processing of *information*. ExamineAprioriEpistemicStatus checks the following factors:

1. Source: If there is no data concerning the source, then the epistemic quality is *information*. A warrant is needed, if *information* will be utilized in later activities.

2. Reliability^p of the information creation process: If there is no data concerning the reliability^p or the reliability^p is below the threshold of belief, then the epistemic quality is *information*. A warrant is needed, if *information* will be utilized in later activities.

3. Warrantor: If there is no data concerning the warrantor, then the epistemic quality is *information*. A warrant is needed, if *information* will be utilized in later activities.

4. Security factors: If the security factors are not at required level, then the epistemic quality is *information* and it should be deleted.

The ObtainWarrant phase consists of searching a warrantor and querying the epistemic quality of information (a warrant). ObtainWarrant creates and activates a SearchWarrantor thread that queries from a warrantor service directory a warrantor that provides warrants in the domain of *information*.[8] If a Warrantor is found then ObtainWarrant sends *information* to the Warrantor for the evaluation of the epistemic quality. After this phase there is a certain understanding of the epistemic quality of *information*, that is the reliability^p of the *information* creation and/or warranting process.

The *utilize – distribute information* phase consists of the following agents and actions (Figure 6.18): ContextEvaluator, DistributionRequisite, and InformationDistributor. ContextEvaluator determines possible worlds and selects relevant worlds. It estimates the severity of each relevant world

[7] For further data, see Appendix *Belief Description Framework*.

[8] SearchWarrantor can also have its own list of needed warrantors.

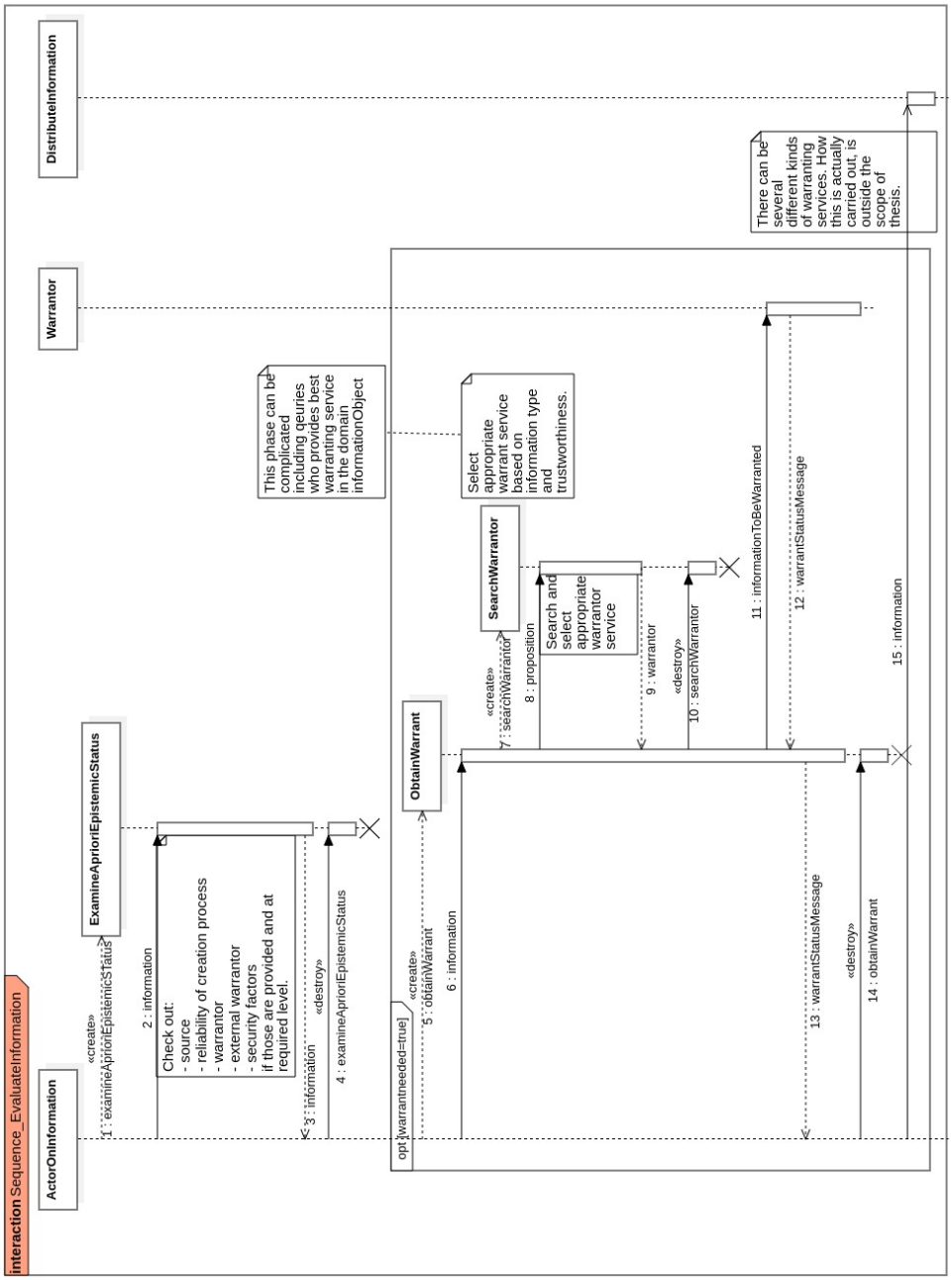

Figure 6.15: BDF sequence diagram of evaluation of information.

(if necessary) and determines the reliability^p thresholds for each epistemic quality of *information*.[9] DistributionRequisite sets the epistemic quality of *information* and evaluates whether to carry out activities on the basis of the epistemic quality of *information*.[10] If a declaration needs to be announced then ActorOnInformation creates and activates InformationDistributor that takes care of transmitting a notice, warning, or alert to a user.

6.4 BDF and DIDS

In the future DIDS will be a dynamic and adaptive service provider, with which individual ISA$_{bdi}$s and groups of ISA$_{bdi}$s (as well as human beings) associate when ever needed and from which dissociate when needed no more. In this kind of environment there are two viewpoints to BDF, which are the viewpoint of an individual ISA$_{bdi}$ and the viewpoint of the infrastructure of DIDS. In the case of the individual ISA$_{bdi}$ BDF provides tools for the ISA$_{bdi}$ to have beliefs, perceive information to form beliefs, determine the epistemic quality of beliefs, and distribute or act on beliefs. From the viewpoint of the individual ISA$_{bdi}$ all other ISA$_{bdi}$s and human beings providing services in DIDS form an infrastructure in which the ISA$_{bdi}$ operates. A basic infrastructure of DIDS should comprise the following services:[11] a service directory of warrant services, an evaluation the reliability^p of information (warrant service), a control of information transfer, security, and behaviour.

6.5 Summary of BDF

We have introduced a framework to describe and manage different epistemic qualities of information. We modelled information as a concept that comprises the structure of various concepts. Each concept in the structure affects the determination of the epistemic quality of information.

BDF constitutes information as a structure, where the central point is a class Information that has relationships (dependency, association, generalization, or realization) with other classes that are Proposition; Presentation; EpistemicQuality comprising EpistemicTheory; Warrant comprising Warrantor;

[9] This phase comprises several unsolved problems dealing with modal logics, frame problem, and omniscience.

[10] This activity is application dependent and therefore, outside the scope of this thesis.

[11] This depends very much on the application domain and its requirements.

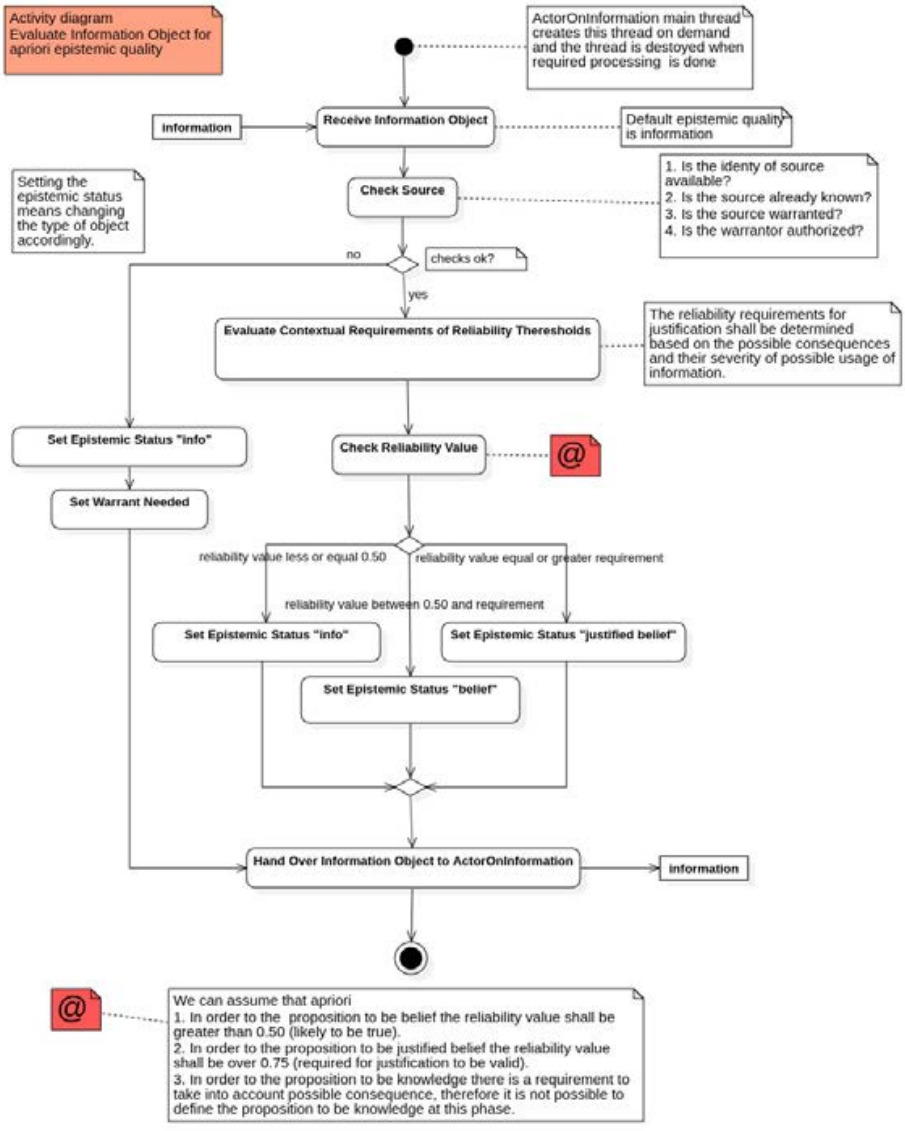

Figure 6.16: BDF activity diagram of evaluation of information – apriori.

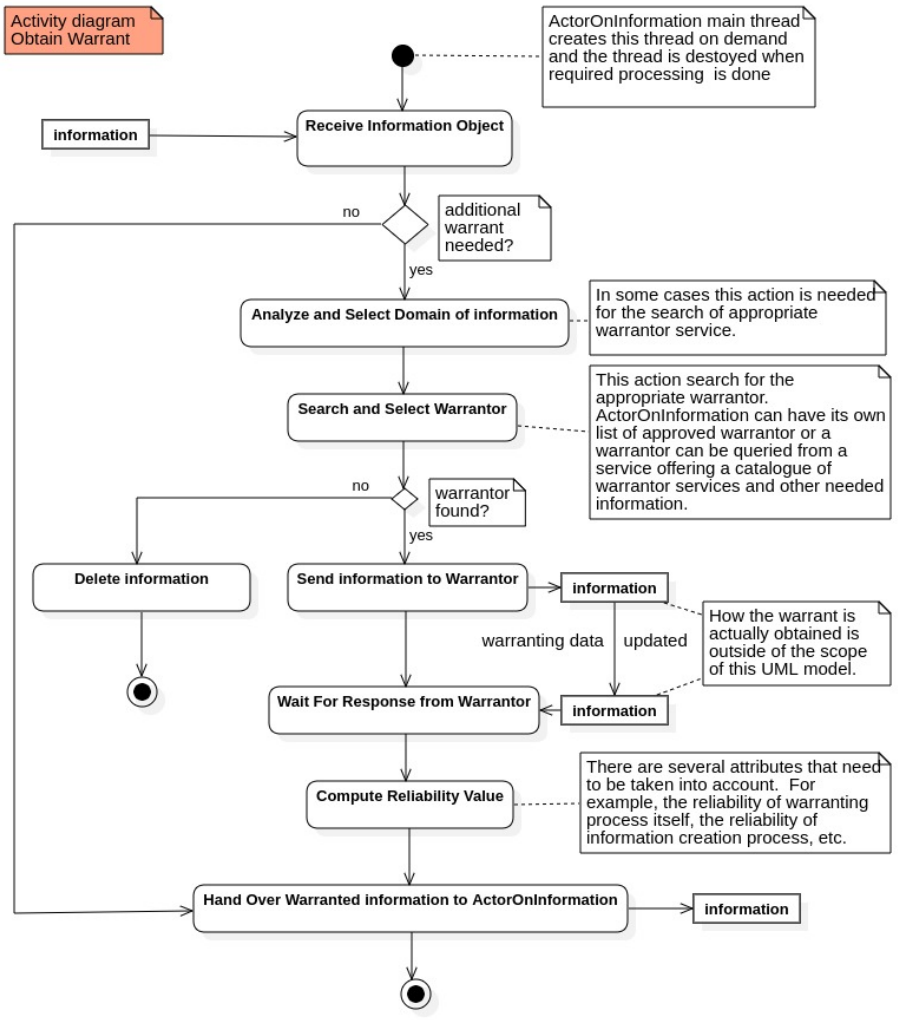

Figure 6.17: BDF activity diagram of evaluation of information – warrant.

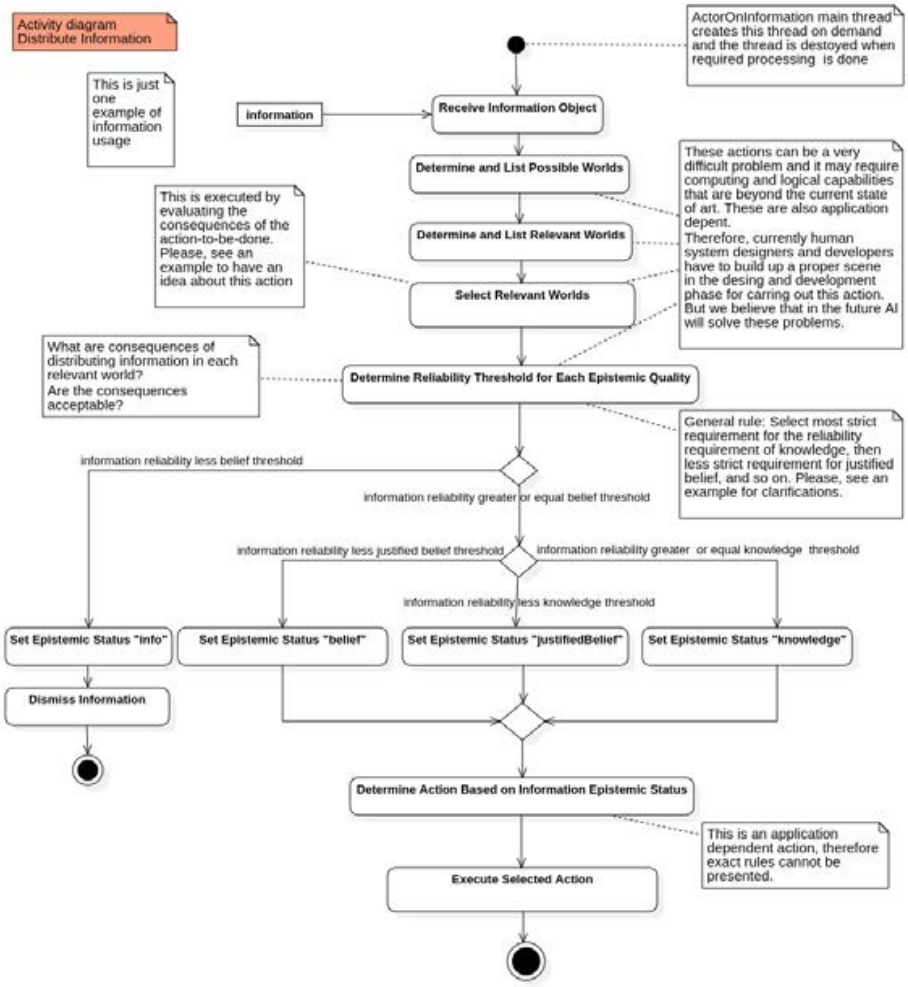

Figure 6.18: BDF activity diagram of distributing information.

Security; Context comprising PossibleWorlds, ActualWorld, and CounterfactualWorld; and ActorOnInformation comprising InformationSource and InformationUser). We see BDF as a part of Virtual Machine Functionality.[12]

We showed how BDF fulfils the requirements of the epistemic theory of Pragmatic Process Reliabilism. We also showed with the help of the TIS scenario some examples of instantiated classes and processing *information*.

We claim that BDF provides users with several advantages. First, BDF is more close to human beings' concept of information and usage of information than current approaches, which mainly implicitly assumes that perceived information is true if it passes a rather rough evaluation. BDF enables explicit specifications of the environment of information and the epistemic quality of information (human being manages quite often unconsciously these issues). Therefore, ISA_{bdi}s and DIDS systems can express in the same sense as human beings phrases such as *"I believe that ..."*, *"I don't believe that ... because ..."*, *"I know that ..."*, and *"I know that ... because ..."*. In addition, they are able motivate their actions to users, for example, by stating *"I cannot do the intended action because I don't know the required requisites well enough."* and *"I don't believe the situation to be beneficial to carry out the task because information is not reliable enough ..."*. This is more close to future social contexts of DIDS where human beings and ISA_{bdi}s co–operate to produce and use services. Second, BDF provides better tools to address dependability issues that are related to the uncertainties of information. It enables to specify the limits of the epistemic quality of information demanded by the expected consequences of actions, so that ISA_{bdi} is better able to plan and execute (or not to execute) actions based on intentions. BDF enables more detailed analyses of faults, failures, and errors in the cases where the epistemic quality of information plays a role. In addition, BDF enables a good base to manage the degraded mode of operation in situations where the epistemic quality of information varies. Third, BDF provides also mechanisms for the information exchange in collaborations to better take care of the epistemic quality of information.

But it is a challenging task to prove that our above claims are valid. It requires a lot of work and a multidisciplinary project that comprises of the expertise in the domains of AI, modal logics, computer science, cognitive science, and epistemology. As we discuss these problems in this thesis at a high, theoretical level, we consider the actual proof to be outside the scope of this thesis; hence, it is the topic of future research.

[12]See page 53.

There are several unsolved problems to implement and utilize BDF fully. The problems deal with the following issues:

- Modal logics do not cover the varying epistemic quality of information;
- Defining and implementing of possible worlds;
- Frame problem; and
- Omniscience.

But we claim that a lot can be implemented using current technology, and in the future many of the problems will be resolved.

Chapter 7

Conclusions

In this thesis we have discussed the role of the epistemic quality of information in the new environment of information services provided by future dependable intelligent distributed systems. The new environment will be created by the developments of AI, intelligent software agents, and robotics and the services provided by them. We defined an intelligent distributed system to be as follows:

An intelligent distributed system is a collection of independent agents that appears to its users as a singe coherent system, where an independent agent can be either an intelligent software agent, a robot, a process running in a computer, or a human being, and some of the independent agents are software–based entities, of which some are implemented utilizing artificial intelligence.

The difference between the definitions of distributed system and intelligent distributed system is that in IDS the independent agent is an intelligent entity (intelligent software agent, robot, or human being) and not just a computer. IDS will process (perceive, create, modify, act on, distribute) information, the epistemic quality of which may vary significantly.

Epistemology is the study of knowledge and justification concerning on what basis human being is justified to believe or does know something. We are of the opinion that in the joint environment of intelligent software agents and human beings we should use, whenever it is possible, concepts that are used by and familiar to human beings. Therefore, we adapted epistemological theories to the dependability theory. We argued that epistemological theories establish a solid ground also for evaluating the epistemic quality of information in the environment of ISA_{bdi} and DIDS. First, we addressed the problem of anthropomorphism and discussed our motivation for intelligent software–based entities to have epistemological concepts, such as belief, justified belief and knowledge. Second, we showed that there is an

important role of justified belief and knowledge in DIDS. Third, we discussed logical issues related to belief and knowledge. We discussed that epistemic logic can be used to solve some of the problems, but there are still several unsolved problems for future research.

One of the main contributions of this thesis is the analysis of the epistemological concepts—belief, justified belief, and knowledge—in the contexts of ISA_{bdi} and DIDS. We introduced the theories of Pragmatic Process Reliabilism that can be adopted by dependable ISA_{bdi}s and DIDS. We defined the concepts as follows:

Belief An epistemic agent has beliefs that have similar features compared to human being's beliefs. Belief is defined as follows:

Belief is a propositional attitude,

1. *which is the state of having an opinion about something to be the case;*
2. *which is created by its actual and potential causal relations to sensory stimulations, behaviour, and/or other propositional attitudes; and*
3. *the representation of which—structured if necessary—is stored in a linguistic form.*

Justified Belief Pragmatic process reliabilism explicates the justification in the joint contexts of ISA_{bdi} and human being. The definition of justification is as follows:

An epistemic agent has justification for her/his/its belief that p if,

1. *The epistemic agent believes p to be true;*
2. *The epistemic agent's belief that p was produced through reliable^p processes P_i; and*
3. *The reliability^p of the processes P_i is adequately high for the requirements set by the contextual factors in the environment where the agent utilizes the belief that p.*

Knowledge Pragmatic process reliabilism explicates well the way we understand the concept of propositional knowledge. Knowledge is defined as follows:

An epistemic agent knows that p if and only if

1. *p is true;*
2. *The epistemic agent believes p to be true;*
3. *If the epistemic agent were to believe that p, p would not be false;*
4. *The epistemic agent's belief that p was produced through reliable^p processes P_i; and*
5. *The reliability^p of the processes P_i either exceeds or is equal to the reliability^p requirements of the actions,*
   (a) *where the agent utilizes the belief that p and*
   (b) *which are set by the expected consequences of the actions.*

In order to cope better with the situations created by the variation of the epistemic quality of information we enhanced the dependability taxonomy with the following concepts:

- Three new attributes:
  1. Skillfulness
  2. Truthfulness
  3. Serveability.
- Two new fault classes:
  1. Training fault
  2. Learning fault.
- Two new service failure concepts:
  1. Action failure
  2. Observed failure.
- Two new means:
  1. Relearning
  2. Retraining.

We are strongly of the opinion that with these new concepts we are able to better formalize and develop methods to manage and improve the dependability of future intelligent distributed systems.

Another main contribution of this thesis is Belief Description Framework. First, we defined relationships between epistemological concepts and software entities (classes). Second, we showed that information, belief, justified belief, and knowledge can be specified as classes and then instantiated as objects. The Information class defines the environment—a kind of information ecosystem—of information. It is the central point. It has relationships with other classes: Proposition, Presentation, EpistemicQuality, Warrant, Security, context, and ActorOnInformation. Third, we specified some important requirements for BDF. Fourth, we showed by modelling BDF using the UML modelling method that BDF can be specified and implemented.

The summary of contributions of this thesis are as follows:

1. We introduced an epistemological approach based on the epistemic quality of information to the dependability of ISA_{bdi} and DIDS. This is the major contribution of this thesis.

2. We provided methods and tools for better understanding dependability issues related to the epistemic quality of information in DIDS. We discussed these issues in Section 2.1.1 *Scenarios*, in Chapters 4 *Belief as Dependability Factor* and 6 *Belief Description Framework*.

3. We carried out careful analyses of epistemic value, truth, trust, and trustworthiness in the joint context of artificial (intelligent software agents) entities and human beings. We discussed these topics in Sections 3.2 *Epistemic Value*, 3.3 *Truth*, and 3.7 *Trust*.

4. We specified enhanced definitions of belief^p, justified belief^p, and knowledge^p to be adapted in the joint context of artificial (intelligent software agents) entities and human beings. We introduced and discussed these definitions in Sections 3.4 *Belief*, 3.5 *Justified Belief*, and 3.6 *Knowledge*.

5. We enhanced the dependability taxonomy to include concepts that can be utilized in the contexts of ISA_{bdi} and DIDS. We introduced these in Chapter 5 *Enhancement to Dependability Taxonomy*.

6. We introduced Belief Description Framework that specifies one proposal to implement and manage the different epistemic qualities

of information. We discussed BDF in Chapter 6 *Belief Description Framework*.

7. We defined a simple UML model to show implementability of Belief Description Framework. We described this in Appendix *Belief Description Framework*.

Future Research Topics

As this thesis is one of the first steps to adapt epistemological theories to future dependable intelligent distributed systems, there are several various topics for future research.

In the domain of epistemology there are many open issues. First, the explication of PPR comprises the anti–Gettier (anti–luck) condition that is difficult both to specify exactly and to implement efficiently. A topic of future research is how to specify nearby worlds and how to implement them without affecting the performance of ISA_{bdi} and DIDS too much. Second, a similar research topic is the evaluation of the required reliability^p of both justified belief and knowledge. Especially, the analysis of possible worlds (actual world and counterfactual worlds) in the context of ISA_{bdi} demands more research. Third, what are the roles and effects of defeaters in the context of ISA_{bdi}?

In the domain of modal logics a topic of future research is the logic of justification. Most contemporary studies of epistemology–related logic concentrate on the logic of belief and knowledge, but there are many open problems of the logic of justification.

In the domain of dependability the main future topic is an empirical research on the effects of the epistemic quality of information on the dependability of ISA_{bdi} and IDS. In the domain of BDF some of the important research topics are an implementation of BDF and a trial to prove the actual benefits of the epistemological approach to the dependability.

References

[1] J. Adler. *Epistemological Problems of Testimony*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition, March 2017. Available on https://plato.stanford.edu/archives/spr2013/entries/testimony-episprob.

[2] AI100 Standing Committee and Study Panel. *One Hundred Years Study on Artificial Intelligence (AI100)*. 2016 report, Stanford University, 2016. Available on https://ai100.stanford.edu/2016-report.

[3] E. Alonso. *Action and Agents*. In *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, 2014. ISBN 978-0-521-87142-6.

[4] E. Amir. *Reasoning and Decision Making*. In *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, 2014. ISBN 978-0-521-87142-6.

[5] G. Antoniou and F. van Harmelen. *A Semantic Web Primer*. The MIT Press, Cambridge, Massachusetts, 2004. ISBN 0-262-010210-3.

[6] K. Arkoudas and S. Bringsjord. *Philosophical Foundations*. In K. Frankish and W. Ramsey, editors, *The Cambridge Handbook of Artificial Intelligence*. Cambridge University Press, 2014. ISBN 978-0-521-87142-6.

[7] S. Artemov. *The Logic of Justification. The Review of Symbolic Logic*, 1(4):477–513, 2008.

[8] S. Artemov and M. Fitting. *Justification Logic*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016. Available on https://plato.stanford.edu/achieves/win2016/entries/logic-justification/.

[9] R. Audi. *The Place of Testimony in the Fabric of Knowledge and Justification. American Philosophical Quarterly*, Volume 34(Number 4):405–422, October 1997.

[10] R. Audi. *Testimony as Social Foundation of Knowledge. Philosophy and Phenomenological Research*, Vol LXXXVII(No. 3):507–531, November 2013.

[11] A. Avizienis, J.-C. Laprie, and B. Randell. *Dependability and Its Threats: A Taxonomy. Building the Information Society*, 156:91–120, 2004. IFIP International Federation for Information Processing.

[12] A. Avizienis, J.-C. Laprie, B. Randell, and C. Landwehr. *Basic Concepts and Taxonomy of Dependable and Secure Computing. IEEE transaction on Dependable and Secure Computing*, Vol 1(No. 1), 2004. Available on IEEE Xplore Digital Library ieeexplore.ieee.org.

[13] J. Baker. *Trust and Rationality.* In E. Sosa, J. Kim, J. Fantl, and M. McGrath, editors, *Epistemology: An Anthology*, Blackwell Philosophy Anthologies, chapter Trust and Rationality, page 807–814. Blackwell Publishing, second edition, 2008. ISBN 978-1-4051-6967-7.

[14] T. Baldwin. *The Identity Theory of Truth. Mind*, C:35–52, 1991. Available on https://doi.org/10.1093/mind/C.397.35.

[15] M. A. Boden. *GOFAI.* In K. Frankish and W. M. Ramsey, editors, *The Cambridge Handbook of Artificial Intelligence.* Cambridge University Press, 2014. ISBN 978-0-521-87142-6.

[16] L. Bonjour. *Can Empirical Knowledge Have a Foundation? American Philosophical Quartely*, 15(1):1–13, January 1978.

[17] L. BonJour. *Internalism and Externalism.* In P. K. Moser, editor, *The Oxford Handbook of Epistemology*, Oxford Handbooks Online Philosophy, chapter Internalism and Externalism. Oxford Handbooks Online, online edition, 2005.

[18] L. BonJour. *Can Empirical Knowledge Have a Foundation?* In E. Sosa, J. Kim, J. Fantl, and M. McGrath, editors, *Epistemology: An Anthology*, Blackwell Philosophy Anthologies, chapter Can Empirical Knowledge Have a Foundation?, page 109–203. Blackwell Publishing, second edition, 2008. ISBN 978-1-4051-6967-7.

[19] L. BonJour. *Externalist Theories of Empirical Knowledge.* In E. Sosa, J. Kim, J. Fantl, and M. McGrath, editors, *Epistemology: An Anthology*, Blackwell Philosophy Anthologies, chapter Externalist Theories of Empirical Knowledge, page 363–378. Blackwell Publishing, second edition, 2008. ISBN 978-1-4051-6967-7.

[20] B. Booch, J. Rumbaush, and I. Jacobsen. *The Unified Modeling Language User Guide.* Addition Wesley, 2nd edition, 2005. ISBN 978-0321267979.

[21] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014. ISBN 978-0-19-967811-2.

[22] N. Bostrom and E. Yudkowsky. *The Ethics of Artificial Intelligence.* In *The Cambridge Handbook of Artificial Intellegence.* Cambridge University Press, 2014. ISBN 978-0-521-87142-6.

[23] M. S. Brady. *Curiosity and the Value of Truth.* In A. Haddock, A. Millar, and D. Pritchard, editors, *Epistemic Value.* Oxford University Press, 2009. ISBN 978-0-19-923118-8.

[24] M. Bratman. *Intention, Plans, and Practical Reason.* Harvard University Press, 1987. ISBN 9780674458185.

[25] T. Button. *The Limits of Realism.* Oxford University Press, first edition, June 2013. ISBN 978-0-19-967217-1.

[26] A. Casali, L. Godo, and C. Sierra. *Graded BDI Models for Agent Architectures.* In J. Leite and P. Torroni, editors, *Computational Logic in Multi-Agent Systems.* Springer, Berlin, Heidelberg, 2004. Available on https://www.fceia.unr.edu.ar/ acasali/publicaciones/climav.pdf.

[27] D. Cole. *The Chinese Room Argument.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, spring 2019 edition, June 2019. Available on https://plato.stanford.edu/archives/spring2019/entries/ chinese-room/.

[28] Contributors of Wikipedia. *Trustworthiness.* In *Wikipedia The Free Encyclopedia.* 2019. Available on http://simple.wikipedia.org/wiki/Trustworthiness.

[29] M. Dacier. *On the resilience of the dependability framework to the intrusion of new security threats. Dependable and Historic Computing*, page 238–250, 2011.

[30] DAML Workgroups. *The DARPA Agent Markup Language.* Technical report, DAML. Available on www.daml.org.

[31] M. David. *The Correspondence Theory of Truth.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, fall 2016 edition, September 2016. Available on https://plato.stanford.edu/archives/fall2016/entries/truth-correspondence/.

[32] G. Dewsbury, I. Sommerville, K. Clarke, and M. Rouncefield. *A Dependability Model for Domestic Systems. Computer Safety, Reliability, and Security: 22nd International Conference*, SAFECOMP 2003:103–115, 2003.

[33] F. Dretske. *Epistemology and Information.* In *Handbook of Philosophy of Science*, page 29–47. Elsevier-North, Holland, 2008. Available on https://philarchive.org/rec/FREEAI-2.

[34] H. N. Duc. *Resource-Bounded Reasoning about Knowledge.* PhD thesis, University of Leipzig, 2001. Available on http://www.informatik.uni-leipzig.de/~duc/papers/diss.pdf.

[35] M. Dummett. *Testimony and Memory.* Oxford Scholarship Online, November 1996 / 2003. ISBN-13: 9780198236214.

[36] P. Evans. *Fake White House bomb report causes brief stock market panic.* Canadian Broadcasting Company, 04 2013. Available on http://www.cbc.ca/news/world/story/2013/04/23/business-ap-twitter.html.

[37] A. Fairweather and L. Zagzebski. *Virtue Epistemology: Essays in Epistemic Virtue and Responsibility.* Oxford University Press, 1 edition, 2001. ISBN 978-0195140774.

[38] FIPA Members. *FIPA Communicative Act Library Specification.* In *FIPA Standards.* FIPA, 2002. Available on http://www.fipa.org/specs/-fipa00037/sc00037J.pdf.

[39] FIPA Members. *Foundation for Intelligent Physical Agent.* Technical report, IEEE–FIPA, 2013. Available on www.fipa.org.

[40] L. Floridi. *Information.* In *The Blackwell Guide to the Philosophy of Computing and Information*, page 40–61. 2004. ISBN 0-631-22918-3.

[41] J. Fodor. *LOT2: The Language Thought Revised.* Oxford University Press, first edition, 2008. ISBN 978-0-19-958801-5.

[42] S. Franklin. *History, Motivation, and Core Themes.* In K. Frankish and W. Ramsey, editors, *The Cambridge Handbook of Artificial Intelligence.* Cambridge University Press, 2014. ISBN 978-0-521-87142-6.

[43] M. J. Frapolli. *The Nature of Truth: An updated approach to the meaning of truth ascriptions.* Springer, 2013. ISBN 978-9400793279.

[44] G. Frege. *Gottlob Frege Posthumous Writings*, chapter Logic (1897), page 1–8. Basil Blackwell, 1979. ISBN 0-631-12835-2.

[45] G. Frege. *The Foundations of Arithmetic*. Basil Blackwell, 1980, Orginal 1884. ISBN 978-0810106055.

[46] O. Freiman. *Towards the Epistemology of the Internet of Things. International Review of Information Ethics*, 22(12), 2014. Available on https://philarchive.org/archive/FRETTE.

[47] K. Frost-Arnold. *Trustworthiness and Truth: The Epistemic Pitfalls of Internet Accountability. Episteme*, 1:63–81, 2014.

[48] R. Fumerton. *Externalism and Skepticism*. In E. Sosa, J. Kim, J. Fantl, and M. McGrath, editors, *Epistemology: An Anthology*, Blackwell Philosophy Anthologies, chapter Externalism and Skepticism, page 394–406. Blackwell Publishing, 2008. ISBN 978-1-4051-6967-7.

[49] R. Fumerton and A. Hasan. *Foundationalist Theories of Epistemic Justification*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, September 2018. Available on https://plato.stanford.edu/archives/fall2018/entries/justep-foundational/.

[50] J. Garson. *Modal Logic*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, September 2018. Available on https://plato.stanford.edu/archives/fall2018/entries/logic-modal/.

[51] R. Gaskin. *The Identity Theory of Truth*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016. Available on https://plato.stanford.edu/archives/win2016/entries/truth-identity.

[52] E. Gettier. *Is Justified True Belief Knowledge? Analysis*, 23:121–123, 1963.

[53] M. Glanzberg. *Truth*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, September 2018. Available on https://plato.stanford.edu/archives/fall2018/entries/truth/.

[54] S. Goldberg. *Testimonial Knowledge in Early Childhood, Revised. Philosophy and Phenomenological Research*, (76):1–36, 2008.

[55] A. Goldman. *What is Justified Belief. Justification and Knowledge*, 1979.

[56] A. Goldman and B. Beddor. *Reliabilism*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2011 edition, March 2011. Available on https://plato.stanford.edu/archives/spr2011/entries/reliabilism/.

[57] A. Goldman and B. Beddor. *Reliabilist Epistemology*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016. Available on https://plato.stanford.edu/archives/win2016/entries/reliabilism/.

[58] A. Goldman and E. Olsson. *Reliabilism and the Value of Knowledge*. In A. Haddock, A. Millar, and D. Pritchard, editors, *Epistemic Value*. Oxford University Press, 2009. ISBN 978-0199231188.

[59] A. I. Goldman. *Internalism Exposed*. In E. Sosa, J. Kim, J. Fantl, and M. McGrath, editors, *Epistemology: An Anthology*, chapter Internalism Exposed, page 379–393. Blackwell Publishing, second edition, 2008. ISBN 978-1-4051-6967-7.

[60] A. I. Goldman. *What is Justified Belief?* In E. Sosa, Jaegwon, J. Fantl, and M. McGrath, editors, *Epistemology: An Anthology*, chapter What is Justified Belief?, page 333–347. Blackwell Publishing, second edition, 2008. ISBN 978-1-4051-6967-7.

[61] V. Goranko and A. Galton. *Temporal Logic.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, fall 2008 edition, September 2008. Available on https://plato.stanford.edu/entries/logic-temporal/.

[62] C. Green. *Epistemology of Testimony.* In J. Matheson, editor, *The Internet Encyclopedia of Philosophy.* 2014. Available on https://www.iep.utm.edu/ep-testi/.

[63] S. R. Grimm. *Knowledge, Practical Interest and Rising Tides.* In J. Greco and D. Henderson, editors, *Epistemic Evaluation: Purposeful Epistemolgy.* Oxford University Press, 2015. Available on https://philarchive.org/archive/GRIQPI.

[64] P. Hacker. *Philosophy and Scienticism: What Cognitive Neuroscience Can and What It Cannot Explain.* Video presentation. Available on http://www.youtube.com/watch?v=jjAFqo67yuU.

[65] R. Hardin. *Trustworthiness. Ethics*, Volume 107, Issue 1, October 1996:26–42, 1996.

[66] R. Hardin. *Trust and Trustworthiness*, volume The Russell Sage Foundation Series on Trust. Russell Sage Foundation, 2002. ISBN 978-087154-342-7.

[67] A. Hazlett. *A Luxury of Understanding: On the Value of True Belief.* Oxford University Press, October 2013. ISBN 978-0-19-967480-0.

[68] R. Heersmink. *A Virtue Epistemology of the Internet: Search Engines, Intellectual Virtues, and Education. Social Epistemology*, 2017. Available on https://philarchive.org/archive/HEEAVE.

[69] B. Hestir. *Aristotle's Conception of Truth: An Alternate View. Journal of the History of Philosophy*, Vol 51(No. 2):193–290, 2013.

[70] J. Hintikka. *Knowledge and Belief: An introduction to the Logic of the Two Notions.* Text in Philosophy. Cornell University Press, May 1962 / 2005. ISBN 978-1904987086.

[71] J. Hyman. *The Road to Larissa. Ratio*, XXIII 4 December 2010(4):393–414, December 2010.

[72] J. Ichikawa and M. Steup. *The Analysis of Knowledge.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, summer 2018 edition, June 2018. Available on https://plato.standord.edu/archives/sum2018/entries/knowledge-analysis.

[73] International Electrotechnical Commission (IEC) Secretariat. *Industrial-process measurement and control - Evaluation of system properties for the purpose of system assesment - Part 5: Assessment of system dependability.* Number Part 5. Dec 1994.

[74] ISO/IEC 13250-2:2006 Information Technology. *SGML Applications - Topic maps.* Technical report, ISO, 2006. Available on https://www.iso.org/standards/38068.html.

[75] P. Jacob. *Intentionality.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, spring 2019 edition, March 2019. Available on https://plato.stanford.edu/archives/spr2019/entries/inten tionality/.

[76] W. James. *Pragmatism A New Name for Some Old Ways of Thinking.* Longmans, Green, and Co. 91 and 98 Fift Avenue, New York, 1907.

[77] S. A. Kenny. *Anthropomorphism vs Humanism.* Georg Henrik von Wright Lecture at University of Helsinki, June 2014. Available on https://www.helsinki.fi/en/unitube/video/c2be9093-1305-4341-8d6b-6bddb40cd0d2.

[78] S. Kripte. *Outline of a Theory of Truth. The Journal of Philosophy*, (19):690–716, 1975.

[79] J. Lackey. *Testimonial Knowledge and Transmission. The Philosophical Quartely*, 49(197):471–490, October 1999. Available on https://www.jstor.org/stable/2660497.

[80] J. Lackey. *Testimonial Knowledge and Transmission.* In E. Sosa, J. Kim, J. Fantl, and M. McGrath, editors, *Epistemology: An Anthology*, Blackwell Philosophy Anthologies, chapter Testimonial Knowledge and Transmission, page 855–867. Blackwell Publishing, 2008. ISBN 798-1-4051-6967-7.

[81] M. Lammenranta. *We Can't Know.* In S. Cowan, editor, *Problems in Philosophy: An Introduction to the Major Debates on Knowledge, Reality, Value, and Government.* Bloomsbury Academic, 2017. Available on http://www.bloomsbury.com/.

[82] J.-C. Laprie, editor. *Dependability: Concepts and Terminology.* Springer-Verlag, 1992. ISBN 978-3-7091-9172-9.

[83] J.-C. Laprie. *Dependable Computing: Concepts, Limits, Challenges.* In *The Twenty-Fifth International Symposium on Fault-Tolerant Computing (Special Issue)*, page 42–54. IEEE, 1995.

[84] C. Legg and C. Hookway. *Pragmatism.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, spring 2019 edition, 2019. Available on https://plato.stanford.edu/archives/spr2019/entries/pragmatism/.

[85] K. Lehrer. *Coherentism.* In J. Dancy, E. Sosa, and M. Steup, editors, *A Companion to Epistemology*, Blackwell Companions to Philosophy, chapter Coherentism, page 278–281. Blackwell Publishing, first edition, 2010. ISBN 978-0631192589.

[86] H. Leitgeb. *What Theories of Truth Should be Like (but CannotBe). In Philosophy Compass 2/2*, Volume 2(Issue 2):276–290, March 2007.

[87] H. Lieberman. *Usable AI Requires Commonsense Knowledge.* In *CHI 2008 Proceedings*, volume CHI 2008. Computer Human Interaction, ACM, April 2008.

[88] D. Marian. *The Correspondence Theory of Truth.* In *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, 2016. Available on https://plato.stanford.edu/archives/fall2016/entries/truth-correspondence.

[89] J. McCarthy. *Ascribing mental qualities to machines.* In *Philosophical Perspectives in Artificial Intelligence.* Humanities Press, 1979.

[90] M. McGarth and D. Frank. *Propositions.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, 2018. Available on https://plato.stanford.edu/archives/spr2018/entries/propositions/.

[91] R. McKenna. *Epistemic Contextualism Defended. Syntese*, 192(2):363–383, 2015. Available on https://philarchive.org/archive/MCKECD.

[92] B. P. McLaughlin. *Computationalism, Connectionism, and Philosophy of Mind.* In *The Blackwell Guide to the Philosophy of Computing and Information*, page 135–151. 2004. ISBN 0-631-22918-3.

[93] C. McLeod. *Trust.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, fall 2015 edition, June 2015. Available on https://plato.stanford.edu/archives/fall2015/entries/trust/.

[94] C. Menzel. *Possible Worlds.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, 2016. Available on https://plato.stanford.edu/archives/win2016/entries/possible-worlds/.

[95] P. Menzies. *Counterfactual Theories of Causation.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, 2014. Available on https://plato.stanford.edu/archives/spr2014/entries/causation-counterfactual/.

[96] D. M. Mittag. *Evidentialism.* In J. Matheson, editor, *Internet Encyclopedia of Philosophy.* 2019. Available on https://www.iep.utm.edu/evident.

[97] E. F. Moore and C. E. Shannon. *Reliable Circuits Using Less Reliable Relays. Journal of the Franklin Institute*, (262):191–208 and 281–297, September 1956.

[98] K. Mulligan and F. Correia. *Facts.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, 2017. Available on https://plato/stanford.edu/archives/win2017/entries/facts/.

[99] I. Niiniluoto. *Informaatio, tieto ja yhteiskunta.* Valtionhallinnon kehittamiskeskus, 1989. ISBN 951-9314-80-6.

[100] E. Olsson. *Coherentist Theories of Epistemic Justification.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017. Available on https://plato.stanford.edu/archives/spr2017/entries/justep-cohenrence/.

[101] OMG UML Group. *Unified Modeling Language.* Technical report, Object Management Group, 2017. Available on https://www.omg.org/spec/UML.

[102] I. S. Organization. *Quality Concepts and Terminology, Part one: Generic Terms and Definitions.* Technical Report ISO/TC 176/SC 1 N 93, ISO/TC 176/SC 1, Feb 1992.

[103] C. S. Peirce. *The Essential Peirce: Selected Philosophical Writings*, volume 1. Indiana University Press, 1992. ISBN 978-0253207210.

[104] N. Petersen, J. L. Linding, and C. Wright. *Pluralist Theories of Truth.* In *The Stanford Encyclopedia of Philosophy.* Metaphysics Re-

search Lab, Stanford University, winter 2018 edition, 2018. Available on https://plato.stanford.edu/archives/win2018/entries/truth-pluralist/.

[105] A. Plantinga. *Warrant and Proper Function*. Oxford Scholarship Online, 1993. ISBN 978 01950 78640.

[106] J. L. Pollock. *Procedural Epistemology - At the Interface of Philosophy and AI*. In J. Greco and E. Sosa, editors, *The Blackwell Guide to Epistemology*. Blackwell Publishing, 1999. ISBN 0-631-20291-9.

[107] D. Pritchard. *Sensitivity and Safety*. In J. Dancy, E. Sosa, and M. Steup, editors, *A Companion to Epistemology*, Blackwell Companions to Philosophy, page 732–736. Blackwell Publishing, second edition edition, 2010.

[108] D. Pritchard. *Epistemic Paternalism and Epistemic Value*. *Philosophical Inquires*, 1(2), 2013.

[109] D. Pritchard. *Veritism and Epistemic Value*. In H. Kornblith and B. McLaughlin, editors, *Alvin Goldman and His Critics*. John Wiley and Sons, Inc., 2016. ISBN 978-0470673676.

[110] D. Pritchard, J. Turri, and J. Adam. *The Value of Knowledge*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2018. Available on https://plato.stanford.edu/archives/spr2018/entries/knowledge-value/.

[111] J. Pust. *Intuition*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, June 2019. Available on https://plato.stanford.edu/archives/sum2019/entries/intuition/.

[112] H. Putnam. *The Project of Artificial Intelligence*. In *Renewing Philosophy*, page 1–18. Havard University Press, 2009. ISBN 9780674760943.

[113] A. Rao and M. Georgeff. *Modeling Rational Agents within a BDI-Architecture*. In *Reading in agents*, page 317–328. Morgan Kaufmann Publishers Inc., 1998. ISBN 1-55860-495-2.

[114] A. Rao and M. P. Georgeff. *BDI Agents: From Theory to Practice*. In *Proceedings of the First Intl. Conference on Multiagent Systems (ICMAS-95)*, 1995.

[115] R. Rendsvig and J. Symons. *Epistemic Logic*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, volume Spring 2019. Metaphysics Research Lab, Stanford University, spring 2019 edition, March 2019. Available on https://plato.stanford.edu/archives/spr2019/entries/logic-epistemic/.

[116] M. Rescorla. *The Computational Theory of Mind*. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition, 2017. Available on https://plato.stanford.edu/archives/spr2017/entries/computational-mind/.

[117] G. Rey. *Searle's Misunderstanding of Functionalism and Strong AI*. In J. Preston and M. Bishop, editors, *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, chapter Searle's Misunderstanding of Functionalism and Strong AI, page 201–225. Oxford University Press, 2002. ISBN 0-19-825057-6.

[118] M. Richard. *When Truth Gives Out*. Oxford University Press, first edition, 2008. ISBN 978-019-958-728-5.

[119] S. Rosenkranz. *The Structure of Justification*. *Mind*, 127:309–338, 2018. Available on https://doi.org/10.1093/mind/fzx057.

[120] S. Ruohomaa. *The Effect of Reputation on Trust Decisions in Inter-Enterprise Collaborations*. PhD thesis, University of Helsinki, 2012. Available on http://urn.fi/URN:ISBN:978-952-10-7912-2.

[121] B. Russell. *The Problems of Philosophy*. Oxford University Press, 1912. ISBN 0-19-888018-9.

[122] G. Sandu. *Logic of Truth*. Lecture Notes, University of Helsinki, 2013.

[123] G. Sandu. *Modal Logic*. Lecture Notes, University of Helsinki, 2014.

[124] T. Sanislav, G. Mois, and L. Miclea. *An Approach to Model Dependability of Cyber-Physical Systems*. *Microprocessors and Microsystems*, 41(C):67–76, 2016.

[125] F. B. Schneider. *Trust in cyberspace*. Technical report, Committee on Information Systems Trustworthiness, 1999.

[126] E. Schwitzgebel. *Belief*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, September. Available on https://plato.stanford.edu/archives/fall2019/entries/belief/.

[127] J. R. Searle. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, 1969. ISBN 0-521-09626-X.

[128] J. R. Searle. *Minds, Brains, and Programs*. *Behavioral and Brain Sciences*, 3(3):417–457, 1980.

[129] J. R. Searle. *Twenty-One Years in the Chinese Room*, chapter Twenty-One Years in the Chinese Room, page 51–69. Claredon Press - Oxford, 2002. ISBN 0-19-925277-7.

[130] J. R. Searle. *Consciousness & the Brain: John Searle at TEDxCERN*. Youtube, May 2013. URL: https://www.youtube.com/ John Searle presentation at TEDxCERN.

[131] K. Segerberg, J.-J. Meyr, and M. Kracht. *The Logic of Action*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, December 2016. Available on https://plato.stanford.edu/archives/win2016/entries/logic-action/.

[132] W. Sellars. *Does Empirical Knowledge Have a Foundation?* In E. Sosa, J. Kim, J. Fantl, and M. McGrath, editors, *Epistemology: An Anthology*, chapter Does Empirical Knowledge Have a Foundation?, page 94–98. Blackwell Publishing, 2008. ISBN 798-1-4051-6967-7.

[133] J. Sindhu. *Coherence-Based Computational Agency*. PhD thesis, Universitat Autonoma de Barcelona, 2010. Available on http://www.iiia.csic.es/files/pdfs/Thesis.pdf.

[134] A. Sloman. *The Computer Revolution in Philosophy*. The Harvester Press, 1978 / 1991 / 2019. Revised online edition available on http://www.bham.ac.uk/research/projects/cogaff/crp/.

[135] A. Sloman. *Epistemology and Artificial Intelligence*. *Expert Systems in the Microelectronic Age, Edinburgh University Press*, 1979. Available on http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-epist-ai.pdf.

[136] A. Sloman. *Alan Turing - His Work and Impact*, chapter Aaron Sloman Draws Together – Virtual Machinery and Evolution of Mind (Part 2), page 574–579. Elsevier, 2013. Available on http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-virt-evo-2.pdf.

[137] A. Sloman and R. Chrisley. *Virtual Machines and Consciousness. Journal of Consciousness Studies (Special issue on Machine Consciousness)*, 10(4-5), 2003. Available on http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-chrisley-jcs.pdf.

[138] M. Smith. *The Logic of Epistemic Justification. Synthese*, page 1–19, 2017. Available on https://doi.org/10.1007/s11229-017-1422-z.

[139] E. Sosa. *Skepticism and Contextualism. Philosophical Issues*, 10:1–18, 2000.

[140] E. Sosa. *A Virtue Epistemology.* Oxford Scholarship Online, 2007. ISBN 9780199297023.

[141] J. Stanley. *Knowledge and Practical Interest.* Oxford University Press, 2005. ISBN 0-19-928803.

[142] M. Steup. *Epistemology.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, winter 2018 edition, December 2018. Available on https://plato.stanford.edu/archives/win2018/entries/epistemology.

[143] D. Stoljar and N. Damnjanovic. *The Deflationary Theory of Truth.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, fall 2014 edition, September 2014. Available on https://plato.stanford.edu/archives/fall2014/entries/truth-deflationary/.

[144] R. Sun. *Connectionism and Neural Networks.* In K. Frankish and W. Ramsey, editors, *The Cambridge Handbook of Artificial Intelligence.* Cambridge University Press, 2014. ISBN 978-0-521-87142-6.

[145] R. Thomason. *Logic and Artificial Intelligence.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, winter 2018 edition, December 2018. Available on https://plato.stanford.edu/archives/win2018/entries/logic-ai/.

[146] A. M. Turing. *Computing Machinery and Intelligence. MIND A Quarterly Review of Psychology and Philosophy*, Vol. Lix. No. 236.:433–460, 1950.

[147] J. Turri, M. Alfano, and J. Greco. *Virtue Epistemology.* In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019. Available on https://plato.stanford.edu/archives/fall2019/entries/epistemology-virtue/.

[148] J. Vogel. *Reliabilism Leveled. The Journal of Philosophy*, 97(1):602–623, 2000.

[149] J. Vogel. *Reliabilism Leveled.* In E. Sosa, J. Kim, J. Fantl, and M. McGrath, editors, *Epistemology: An Anthology*, chapter Reliabilism Leveled. Blackwell Publishing, 2008. ISBN 978-1-4051-6967-7.

[150] G. H. von Wright. *An Essay in Modal Logic. Amsterdam: North-Holland Pub. Co.*, 1951. ASIN: B00181MC11.

[151] W3C RDF Working Group. *RDF 1.1 Semantics*. Technical report, W3C, 2014. Available on https://www.w3c.org/TR/2014/REC-rdf11-mt-20140225/.

[152] W3C RDF Working Group. *RDF Schema 1.1*. Technical report, W3C, 2015. Available on https://www.w3c.org/TR/2014/REC-rdf-schema-20140225.

[153] W3C RDF Working Groups. *Resource Description Framework Semantic 1.1: Concepts and Abstract Framework*. Technical report, W3C. Available on https://www.w3c.org/TR/2014/REC-rdf11-concepts-2014225/.

[154] W3C Semantic Web Working Group. *OWL 2 Web Ontology Language Profiles (Second Edition)*. Technical report, W3C, 2012. Available on https://www.w3c.org/TR/2012/REC-owl2-profiles-20121211/.

[155] W3C Semantic Web Working Group. *Semantic Web*. Technical report, W3C, 2017. Available on https://www.w3c.org/standards/semanticweb/.

[156] W3C Spatial Data on the Web Working Group. *Time Ontology in OWL*. Technical report, W3C, 2017. Available on https://www.w3c.org/TR/2017/CR-owl-time-20170606.

[157] W3C Standard Organization. Technical report. Available on https://www.w3.org/.

[158] W3C Web Ontology Working Group. *OWL 2 Web Ontology Language Document Overview (Second Edition)*. Technical report, W3C, 2012. Available on https://www.w3.org/TR/2012/REC-owl2-overview-20121211.

[159] W3C Web Ontology Working Group. *OWL 2 Web Ontology Language Primer (Second Edition)*. Technical report, W3C, 2012. Available on https://www.w3c.org/TR/2012/REC-owl2-primer-20121211.

[160] W3C XML Working Groups. *Extensible Markup Language (XML) 1.1*. Technical report, W3C, 2008. Available on https://www.w3c.org/TR/2008/REC-xml-20081126.

[161] W3C XML Working Groups. *XML Essentials*. Technical report, W3C, 2017. Available on https://www.w3c.org/standards/xml/core.

[162] G. Weiss. *Multiagent Systems, A Modern Approach to Distributed Artificial Intelligence*. The MIT Press, Cambridge, Massachusetts, 1999. ISBN 0-262-23203-0.

[163] G. Wheeler and L. M. Pereira. *Epistemology and Artificial Intelligence. Journal of Applied Logic*, 2:469–493, 2004.

[164] T. Williamson. *Knowledge and its Limits*. Oxford University Press, 2000. ISBN 978-0-19-925656-3.

[165] M. Wooldridge. *Reasoning about Rational Agents*. The MIT Press, Cambridge, Massachusetts, 2000. ISBN 0-262-23213-8.

[166] M. Wooldridge and N. R. Jennings. *Intelligent Agents: Theory and Practice. Knowledge Engineering Review*, Vol 10, No.2:115–152, 1995.

[167] J. O. Young. *The Coherence Theory of Truth*. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2016. Available on https://plato.stanford.edu/archives/win2016/entries/truth-coherence/.

[168] L. Zagzebski. *From Reliabilism to Virtue Epistemology*. In *The Proceedings of the Twentieth World Congress of Philosophy*, page 173–179. 2000. ISBN 978-1-889680-19-4.

[169] L. Zagzebski. *Linda Zagzebski*. In J. Dancy, E. Sosa, and M. Steup, editors, *A Companion to Epistemology*, page 210–215. Wiley-Blackwell, 2010. ISBN 978-1-4051-3900-7.

Appendices

APPENDIX 1
—
TERMINOLOGY

agent

An entity (either human being or intelligent software agent) that has an important effect on a situation.

anthropomorphism

Anthropomorphism is the attribution of human traits, emotions, or intentions to non–human entities.

availability (computer science)

Availability means the probability that a system is operational at a given time, i.e. the amount of time a device is actually operating as the percentage of total time it should be operating.

belief (computer science)

Beliefs represent the informational state of the agent, in other words its beliefs about the world (including itself and other agents). Beliefs can also include inference rules, allowing forward chaining to lead to new beliefs. Using the term belief rather than knowledge recognizes that what an agent believes may not necessarily be true.

belief (epistemology)

Belief is a propositional attitude,

1. which is the state of having an opinion about something to be the case;
2. which is created by its actual and potential causal relations to sensory stimulations, behaviour, and/or other propositional attitudes; and
3. the representation of which—structured if necessary—is stored in a linguistic form.

bdi architecture (computer science)

The BDI (Belief—Desire—Intention) architecture model implements the principal aspects of Michael Bratman's theory of human practical reasoning, where a rational agent has certain mental attitudes of belief, desire and intention, representing, respectively, the information, motivational, and deliberative states of the agent.

cognition (philosophy)

Cognition is the mental process of acquiring knowledge and understanding through thought, experience, and the senses. It comprises terms such as knowledge, attention, memory and working memory, evaluation, reasoning and computation, problem solving and decision making. Human cognition is conscious and unconscious, concrete or abstract, as well as intuitive and conceptual. Cognitive processes use existing beliefs and knowledge and generate new beliefs and knowledge.

coherentism (epistemology)
Coherentism defines that every justified belief receives its justification from other beliefs in its epistemic neighbourhood.

confidentiality (computer science)
Confidentiality is ensuring that information is not accessed by unauthorized persons.

connectionism (computer science)
Connectionism is a set of approaches in the fields of artificial intelligence that models mental or behavioural phenomena as the emergent processes of interconnected networks of simple units. There are many forms of connectionism, but the most common forms use neural network models.

correspondence theory (epistemology)
Correspondence theory of truth defines that p is true if and only if p corresponds to some state of affairs that exists, and p is false if and only if p corresponds to some state of affairs that does not obtain.

dependability (computer science)
Dependability is the extent to which a critical system is trusted by its users.

desire (computer science)
Desires represent the motivational state of the agent. They represent objectives or situations that the agent would like to accomplish or bring about.

epistemic agent
An entity (either human being or intelligent software agent) that has an important effect on a situation and perceives, holds, processes, and distributes semantic information.

epistemic logic (philosophy)
Epistemic Logic is the logic of knowledge, belief, and justification.

epistemology (philosophy)
Epistemology is the study of knowledge and justified belief.

error (computer science)
Error is that part of the system state that may cause a subsequent failure. Before an error is detected, it is latent. The detection of an error is indicated at the service interface by an error message.

fact / factual (philosophy)
Facts are the objects of certain mental states and acts, they make truth-bearers true and correspond to truths, and they are part of the furniture of the world. Factual is based on or containing facts.

failure (computer science)
Failure of a system is an event that corresponds to a transition from correct service to incorrect service. It occurs when an error reaches its service interface.

fallibilism (philosophy)
No beliefs are so well justified or supported by good evidence that they could not be false. Thus, there is no conclusive justification or non rational certainly for any of our beliefs.

fault (computer science)
Fault is the (adjudged or hypothetical) cause of error. When it produces an error, it is active, otherwise it is dormant.

fault tolerance (computer science)
Fault tolerance is an approach by which reliability of a computer system can be increased beyond what can be achieved by traditional methods. A system can provide its services even in the presence of faults.

fault tolerant system (computer science)
A system is fault tolerant if it can mask the presence of faults in the system by using redundancy. The goal of fault tolerance is to avoid system failure, even if faults are present.

fault tolerant service (computer science)
A fault tolerant service always guarantees strictly correct behaviour despite a certain number and type of faults.

foundationalism (epistemology)
Foundationalism is a view about the structure of justification or knowledge. The thesis is that all knowledge and justified belief rest ultimately on a foundation of non–inferential knowledge or justified belief.

information (general)
Information is that which informs.

information (computer science)
Information is taken as an ordered sequence of symbols from an alphabet. Shannon information: the entropy, H, of a discrete random variable X is a measure of the amount of uncertainty associated with the value of X.

information (philosophy)
In philosophy information is a complex concept, the definition of which cannot be briefly expressed.
See https://plato.stanford.edu/entries/information/

integrity (computer science)
Integrity is the absence of improper system alterations.

intelligent software agent (computer science)

Intelligent Software Agent (ISA) is a computational entity that can be viewed as perceiving and acting upon its environment and that is autonomous in that its behaviour at least partially depends on its own experience.

intelligent distributed system (computer science)

An intelligent distributed system is a collection of independent agents that appears to its users as a singe coherent system, where an independent agent can be either an intelligent software agent, a robot, a process running in a computer, or a human being, and some of the independent agents are software–based entities, some of which are implemented utilizing artificial intelligence.

intention (computer science)

Intentions represent the deliberative state of the agent, that is, what the agent has chosen to do. Intentions are desires to which the agent has to some extent committed.

intention (philosophy)

There are several different kinds of intention defined in philosophy: intention as doing, intention in action, intention as plan, intention related to belief. For further information, see https://plato.stanford.edu/entries/intention/

justified belief (epistemology)

Justified belief is belief for which an agent has a proper support for its truthfulness. Pragmatic Process Reliabilism: an agent has justification for its belief *that p* if,

1. The agent believes *p* to be true;
2. The agent's belief *that p* was produced through reliable$^{\mathrm{p}}$ processes $\mathbf{P}_{\mathrm{i}}$; and
3. The reliability$^{\mathrm{p}}$ of the processes $\mathbf{P}_{\mathrm{i}}$ is adequately high for the requirements set by the contextual factors in the environment where the agent utilizes the belief *that p*.

knowledge (epistemology)

There are several different definitions of knowledge. In this thesis we use the theory of pragmatic process reliabilism which defines knowing as follows: an agent knows *that p* if and only if

1. *p* is true;
2. The agent believes *p* to be true;
3. If the epistemic agent were to believe *that p*, *p* would not be false;
4. The agent's belief *that p* was produced through reliable$^{\mathrm{p}}$ processes $\mathbf{P}_{\mathrm{i}}$; and

5. The reliability^{p} of the processes $\mathbf{P}_i$ either exceeds or is equal to the reliability^{p} requirements of the actions,
   a) where the agent utilizes the belief *that p* and
   b) which are set by the expected consequences of the actions.

maintainability (computer science)
Maintainability is the simplicity and speed with which a system can be repaired or maintained. It is also defined as the ability to undergo modifications and repairs.

modal logic (logic)
Modal logic studies reasoning that involves the use of the expressions *necessarily* and *possibly*. The term *modal logic* is used also more broadly to cover a family of logics with similar rules and a variety of different symbols, such as deontic logic, temporal logic, epistemic logic, and doxastic logic.

ontology (philosophy)
Ontology is the study of the nature of existence, which is concerned with identifying the kinds of things that actually exist, and how to describe them.

ontology (computer science)
Ontology is an explicit and formal specification of a conceptualization.

proposition (computer science)
A proposition is a statement that is either true or false.

proposition (philosophy)
A proposition is the shareable object of an attitude and primary bearer of truth and falsity.

reliabilism (epistemology)
Reliabilism explains important epistemic concepts in terms of the truth-conduciveness of an epistemic agent's reasoning, belief–forming processes, methods, faculties, etc. The epistemic agent's truth–conduciveness is its likelihood to produce true beliefs, thus to avoid false beliefs. The fundamental idea is that belief *that p* is justified on the basis of a reason, or ground, *r* just in case *r* is a reliable indication *that p* is true .

reliability (computer science)
Reliability is defined as the continuity of the correct service.

reliability (philosophy)
Reliability is defined as the probability that a system will produce correct outputs up to some given time t.

safety (computer science)

Safety is defined as the absence of catastrophic consequences on the users and the environment.

safety (philosophy)

Safety is explicated as follows: If an agent were to believe *that p*, *p* would not be false. In other words, in all nearby worlds where the agent believes *that p*, *p* is not false.

semantic information (computer science)

Information is taken as an ordered sequence of symbols from an alphabet that is meaningful in its context of use. Semantic information is defined as well-formed, meaningful and truthful data.

semantic web (computer science)

Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C).

serveability (computer science)

Serveability is the ability to provide a correct service in the environment of uncertain, conflicting, (or even contradictory) information. This expresses a kind of sensitivity to changing epistemic levels of beliefs.

trust and trustworthiness (philosophy)

There are three kinds of trust. In the first kind, people trust other people because they cannot check all the bases for establishing a belief. The formed belief is not resistant to counter–evidence. The second kind of trust involves more than people's willingness to accept other people or to assume things on trust. People may judge an individual or a thing on the basis of a non–ordinary belief. The formed belief is not resistant to counter–evidence. The third kind of trust is the case in which people think it to be rational to hold a belief even though there is a counter–evidence. Trusting requires that a trustor can tolerate some level of risk or vulnerability, at least, to the failure by a trustee to do or to be what the trustor depends on the trustee. Thus, trusting requires that a trustor can 1) be vulnerable to the trustee, 2) think well of the other, at least in certain domains, and 3) be optimistic that the trustee is competent in certain respect. For trust to be warranted (i.e. well–grounded), both parties must be trustworthy.

trustworthiness (computer science)

Trustworthiness of distributed systems asserts that the system does what is required despite environmental disruption, human user and operator errors, and attacks by hostile parties and that it does not do other things. Design and implementation errors must be avoided, eliminated, or somehow tolerated. Addressing only some aspects of the problem is not sufficient. Moreover, achieving trustworthiness

requires more than just assembling components that are themselves trustworthy.

Trustworthiness is assurance that a system deserves to be trusted; it will perform as expected despite environmental disruptions, human and operator error, hostile attacks, and design and implementation errors. A trustworthy system reinforces the belief that will continue to produce expected behaviour and will not be susceptible to subversion.

virtual machine functionality (philosophy)

According to virtual machine functionalism the human mind is one kind of virtual machine, which is operated by a human body. An intelligent software agent is another kind of virtual machine, which is operated by a computer. The basic idea of functionalism is that the essence of a mental state is not to be found in the biology of the brain but rather in the role that plays in one's mind and in the causal relations that it bears to stimuli. Thus, functionalism claims that mental states are not only physical states, but also functions or operation of those physical states.

APPENDIX 2

—

UML MODEL OF BELIEF DESCRIPTION FRAMEWORK

This is an example of a high level UML model describing our ideas about Belief Description Framework (BDF). The diagrams consist of the following diagrams:

1. Profile diagrams, which define concepts and stereotypes and their relationships.
   Figures 1, 2, 3, 4, 5, 6, and 7.

2. Class diagrams, which define important base classes of BDF.
   Figures 8, 9, 10, 11, 12, 13, 14, and 15.

3. Object diagrams, which provide one example of possible implementation of BDF.
   Figures 16, 17, 18, and 19.

4. Use case diagrams, which define some examples of possible use cases of BDF.
   Figures 20, 21, 22, and 23.

5. Interaction diagrams, which provide examples of possible flows of messages and control.
   Figures 24, 25, and 26.

6. Activity diagrams, which provide some examples of the flow of control and algorithms.
   Figures 27, 28, 29, and 30.

7. An use case diagram of possible worlds, which illustrate an evaluation of the requirements of reliability.
   Figure 31.

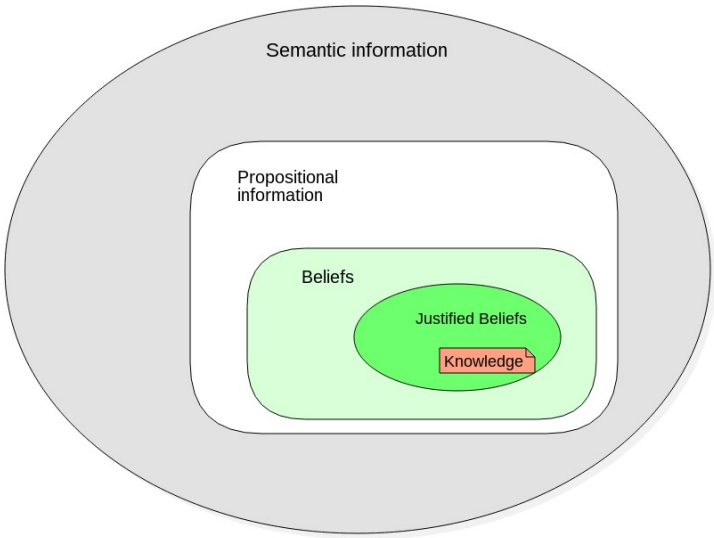

Figure 1: Classification of propositional information.

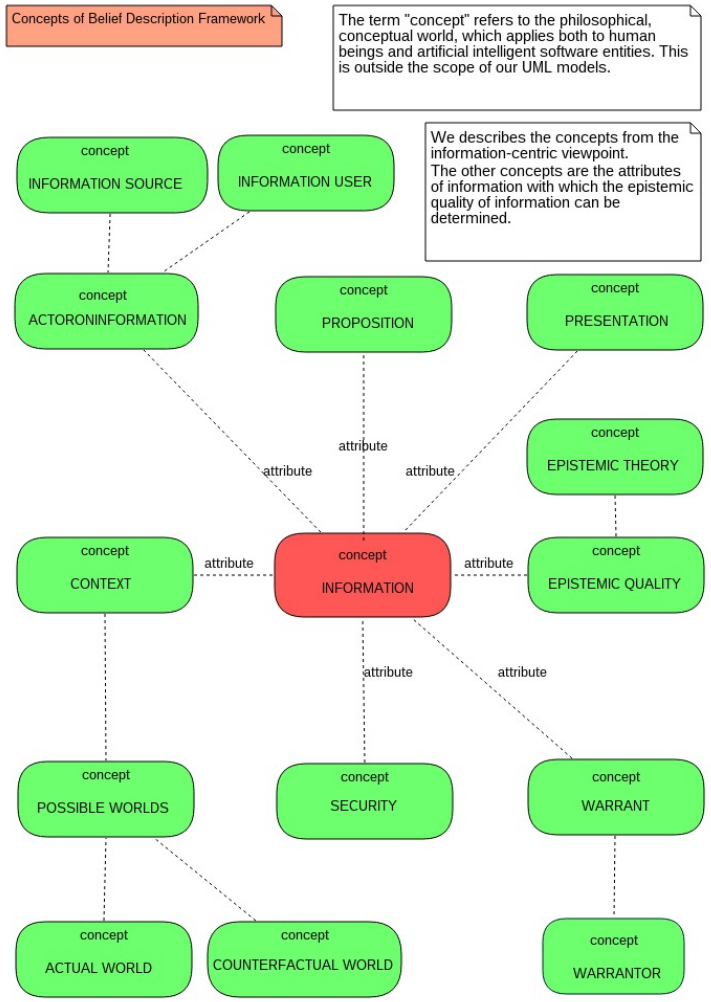

Figure 2: Concepts of information.

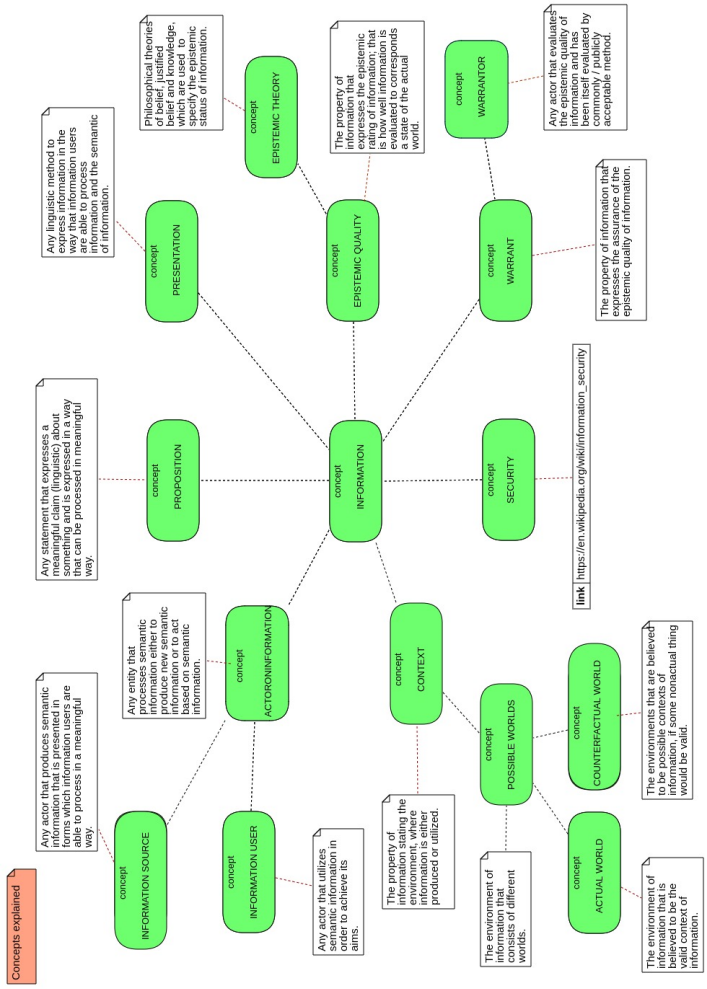

Figure 3: Concepts of information explained.

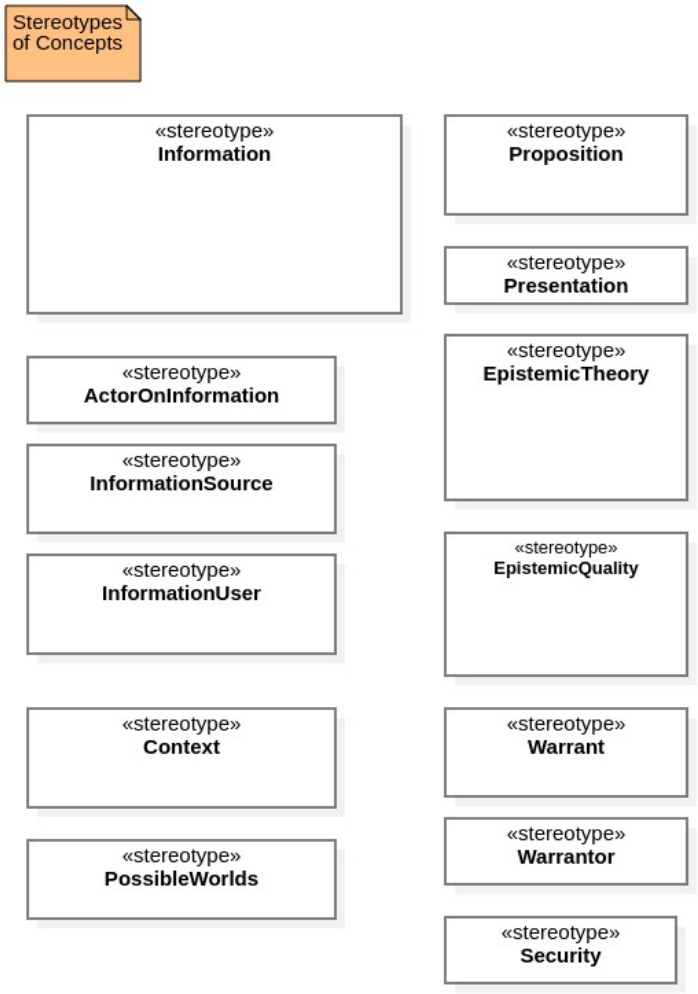

Figure 4: Stereotypes of concepts.

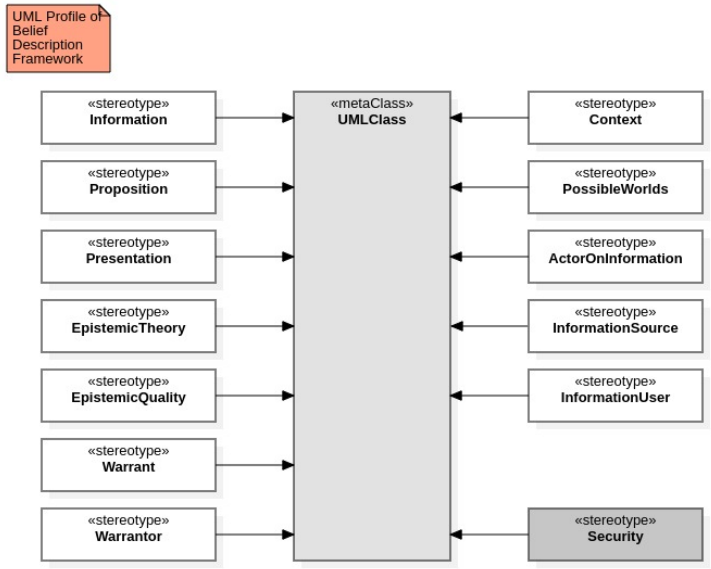

Figure 5: UML profile of Belief Description Framework.

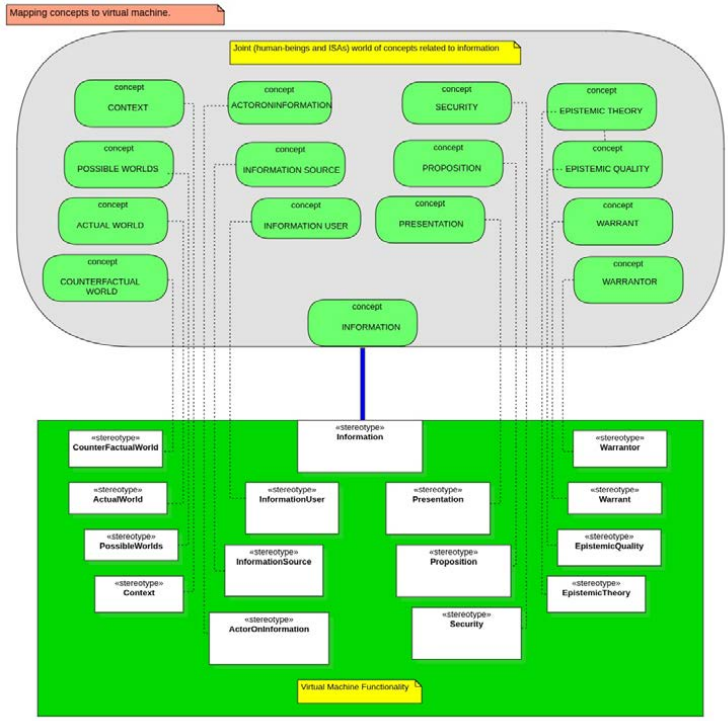

Figure 6: Mapping concepts to virtual machine.

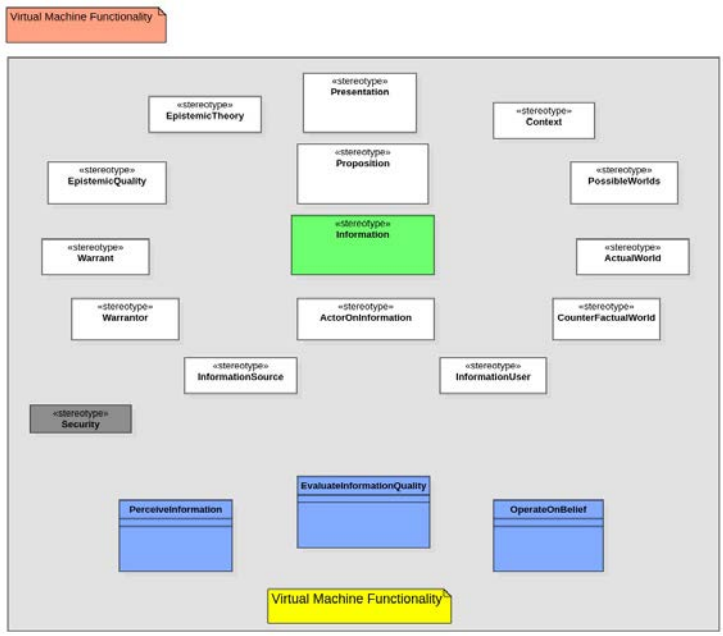

Figure 7: Virtual machine functionality.

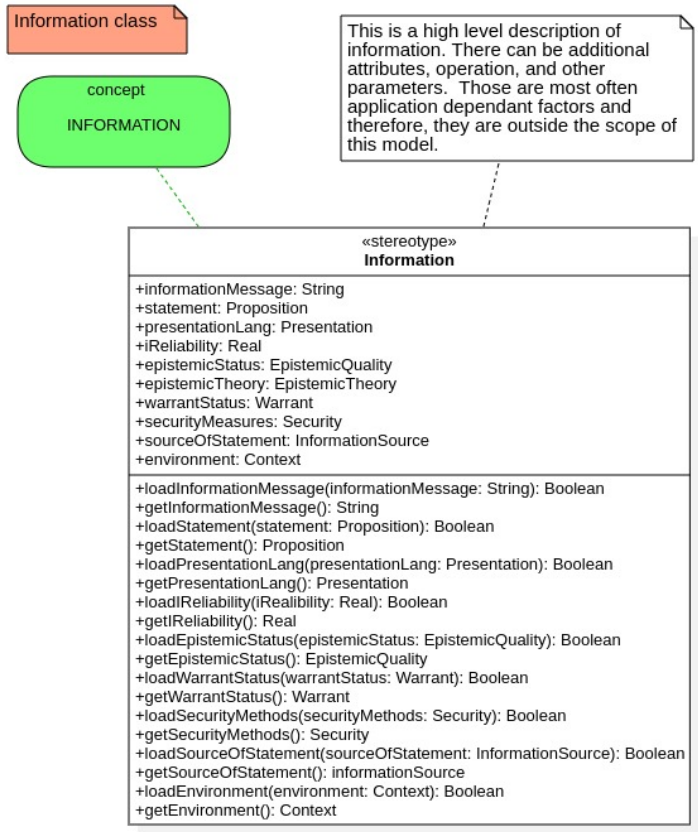

Figure 8: Class diagram of information.

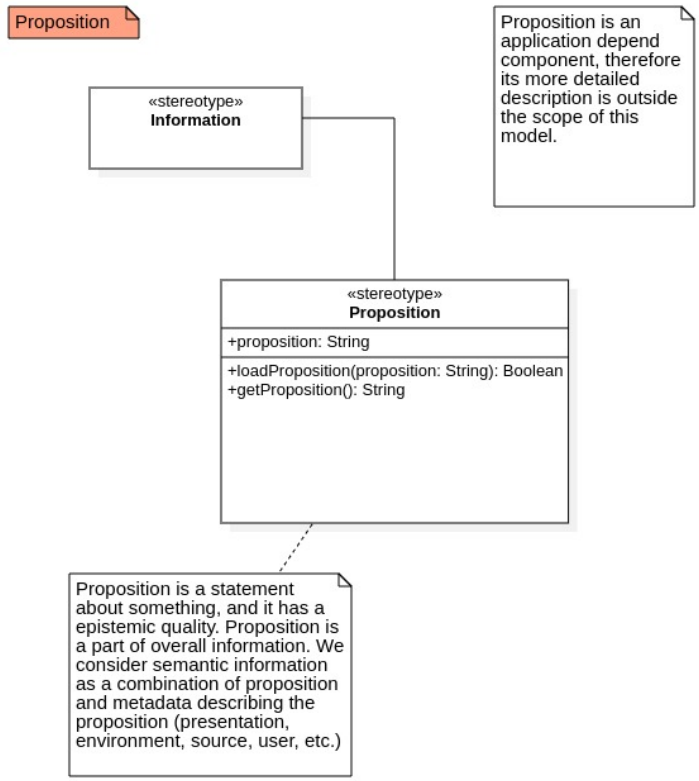

Figure 9: Class diagram of proposition.

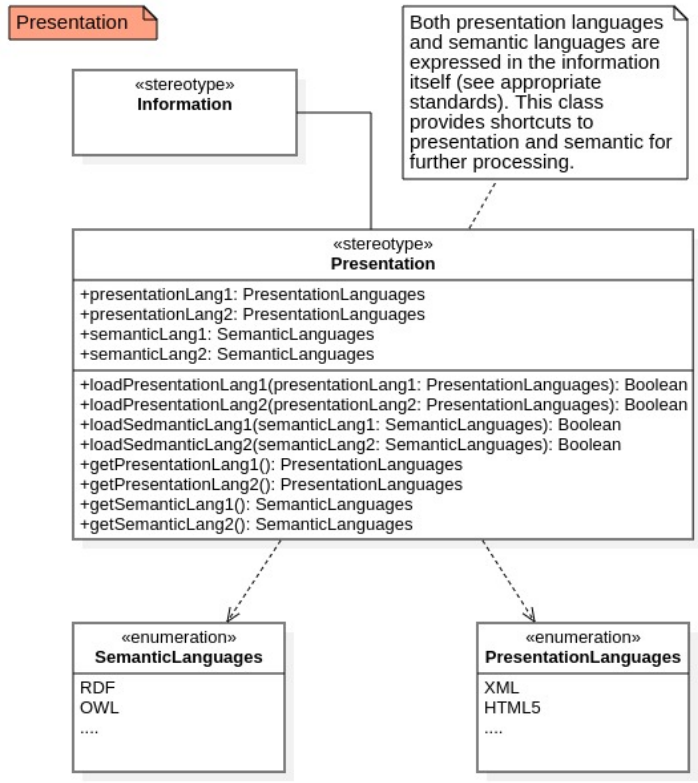

Figure 10: Class diagram of presentation.

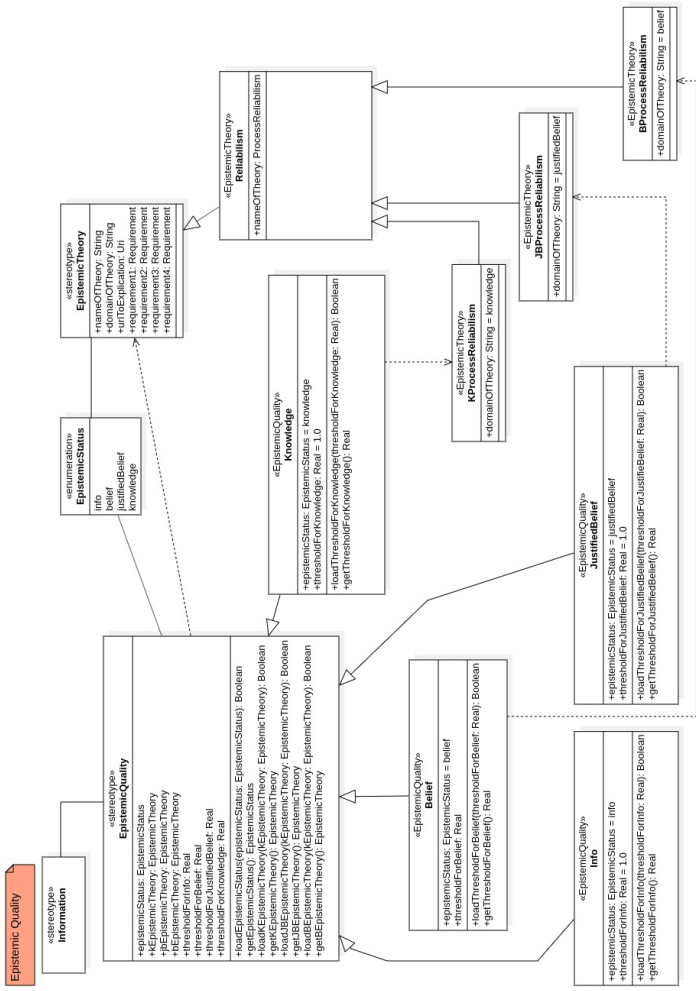

Figure 11: Class diagram of epistemic quality.

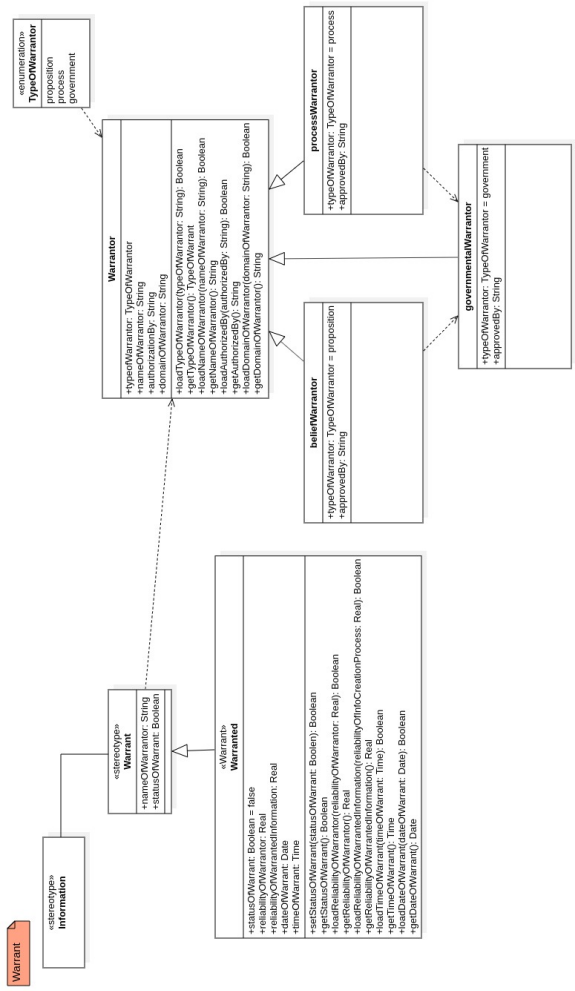

Figure 12: Class diagram of warrant and warrantor.

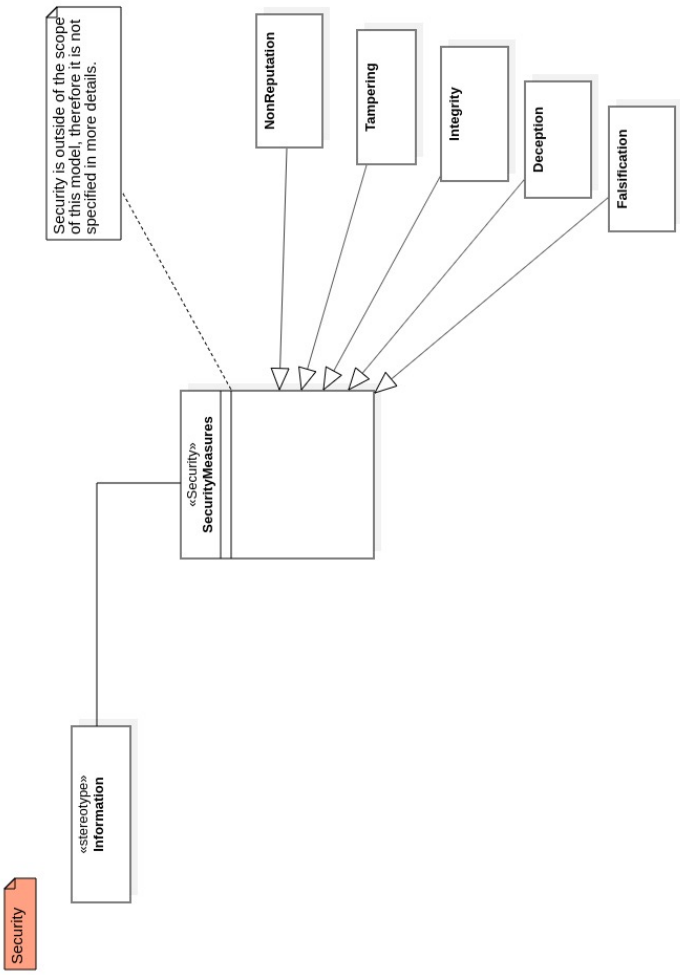

Figure 13: Class diagram of security.

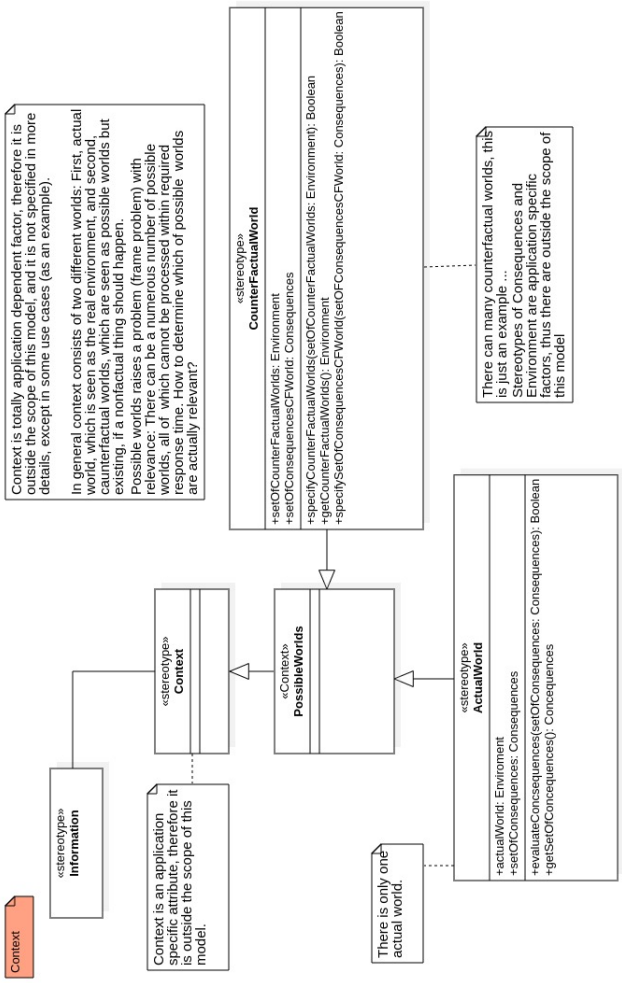

Figure 14: Class diagram of context.

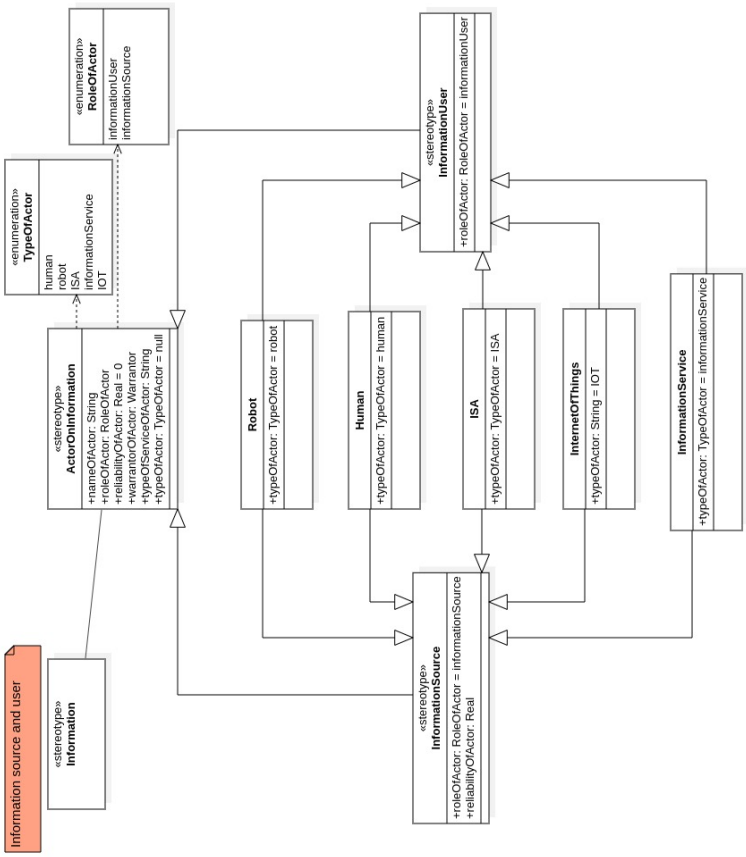

Figure 15: Class diagram of information source and user.

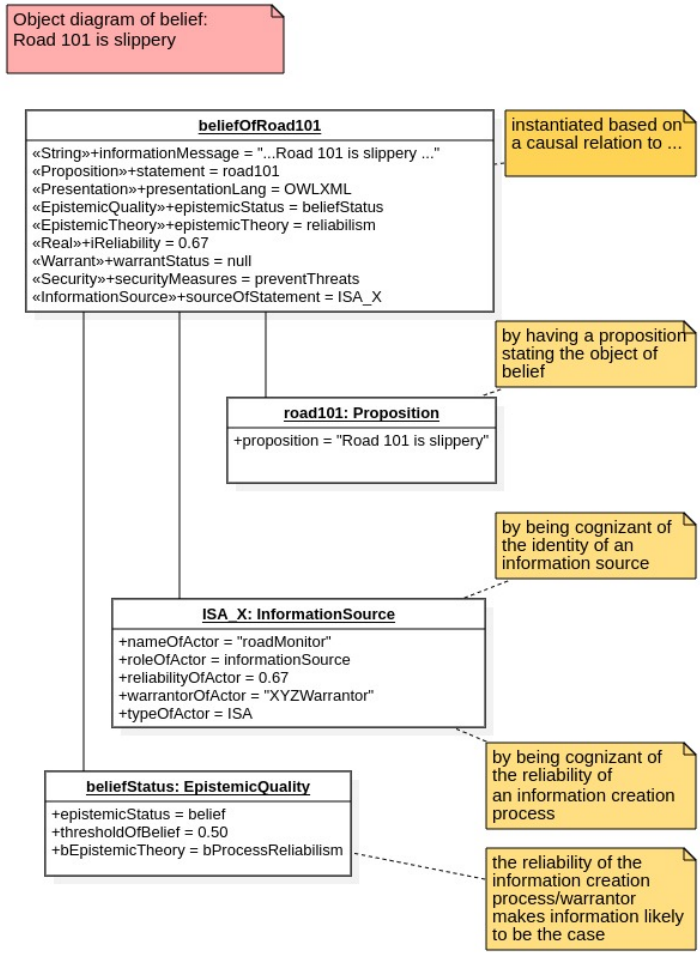

Figure 16: Object diagram of belief.

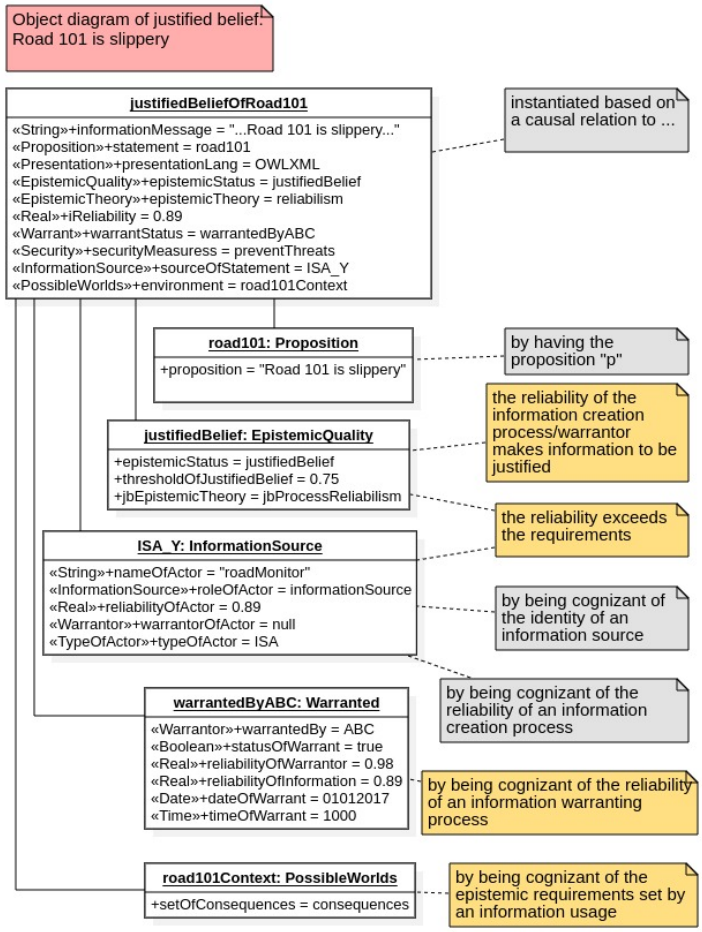

Figure 17: Object diagram of justified belief.

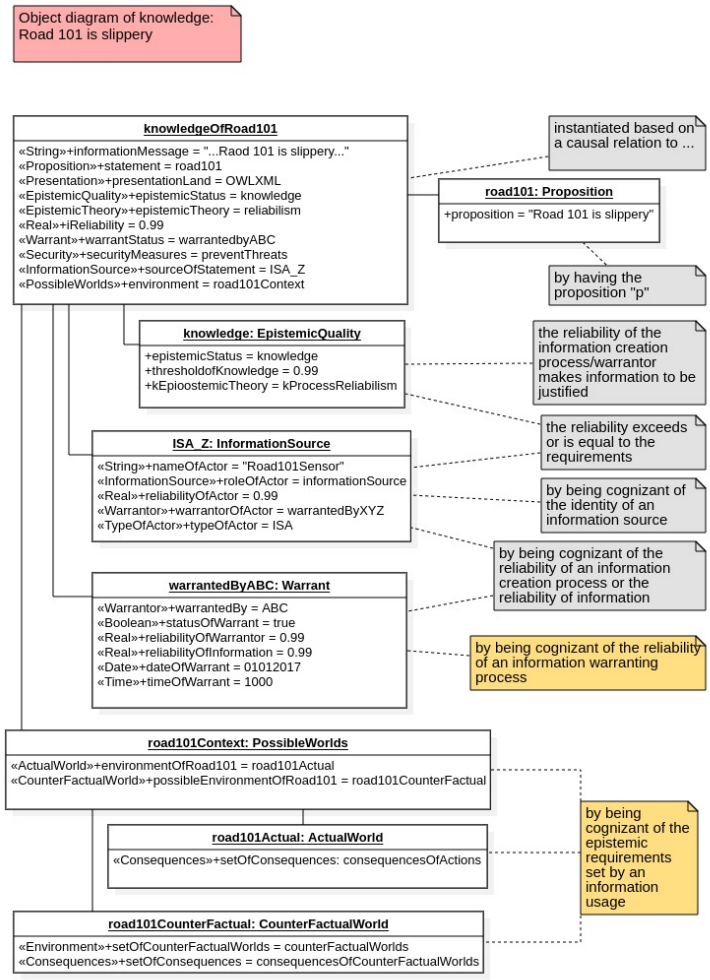

Figure 18: Object diagram of knowledge 1.

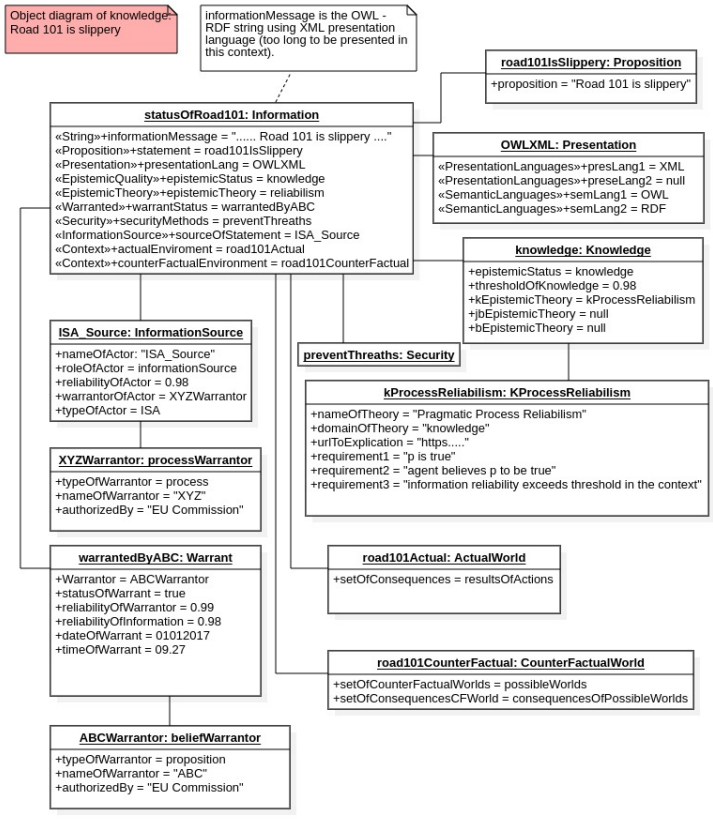

Figure 19: Object diagram of knowledge 2.

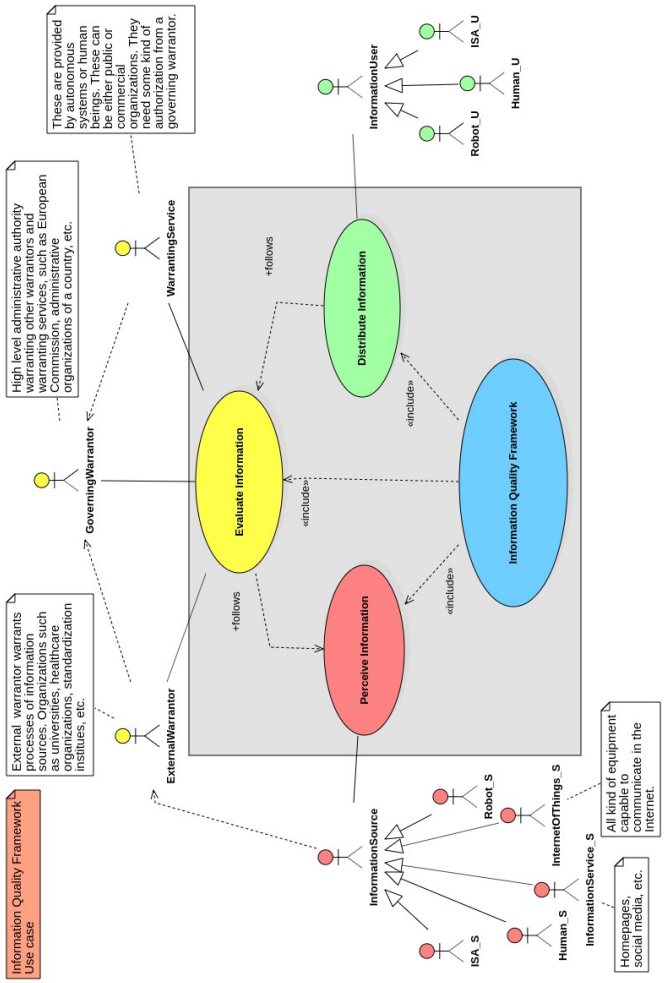

Figure 20: Overall use case diagram of Belief Description Framework.

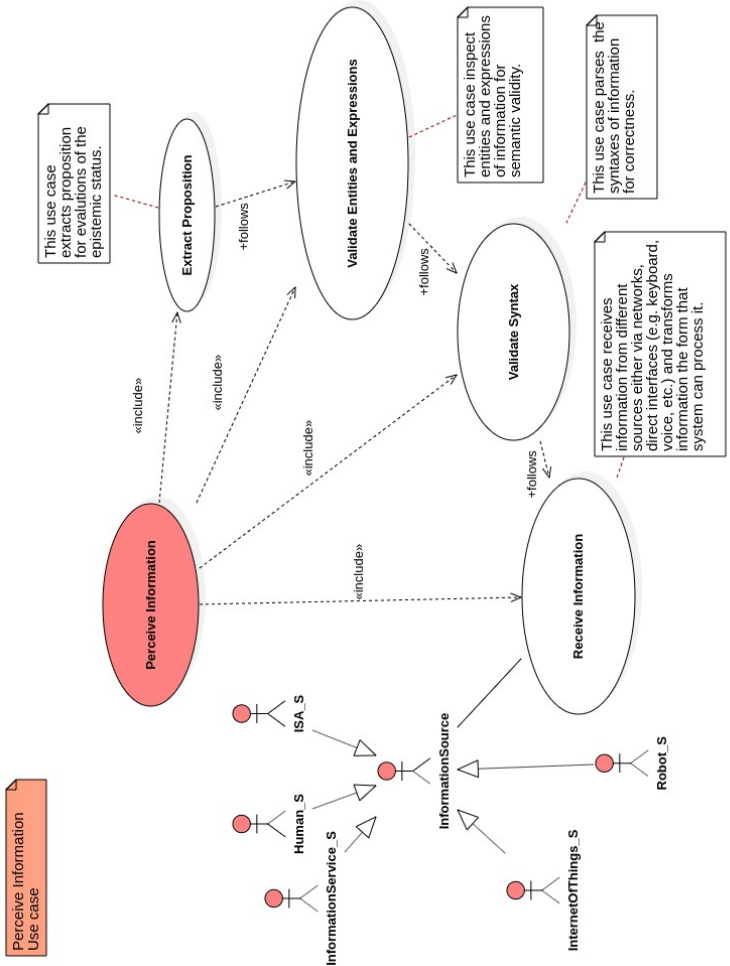

Figure 21: Use case diagram of perceive information.

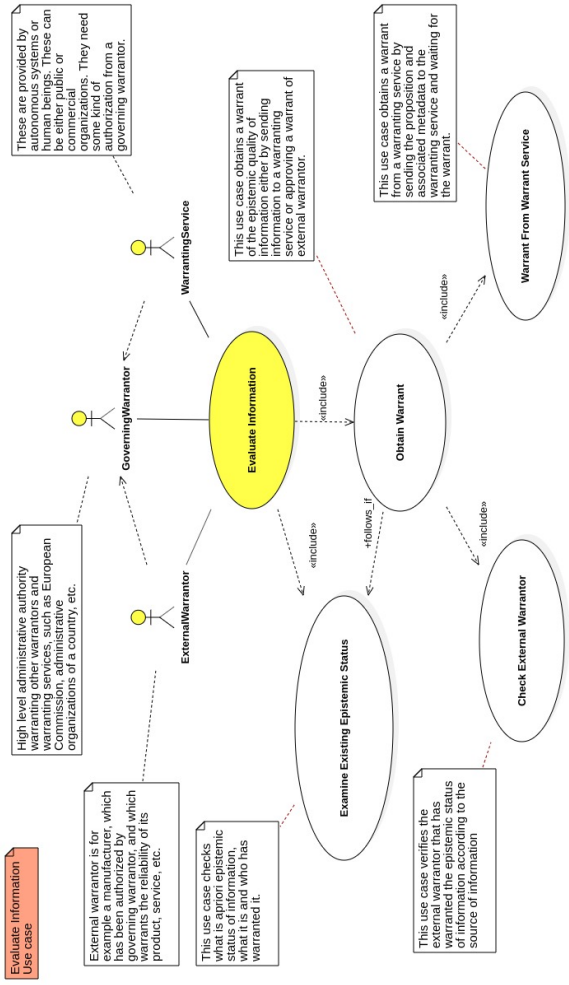

Figure 22: Use case diagram of evaluate information.

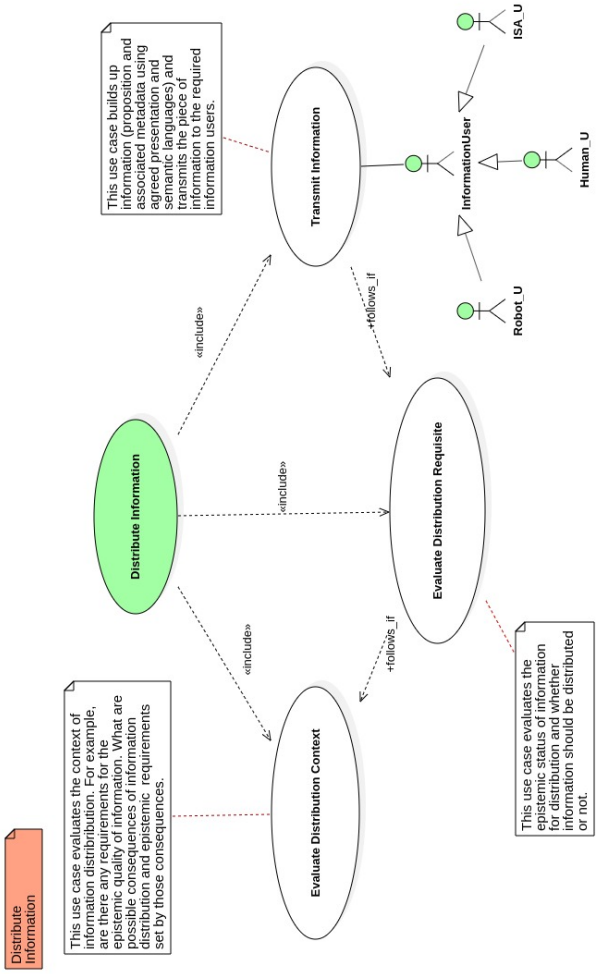

Figure 23: Use case diagram of distribute information.

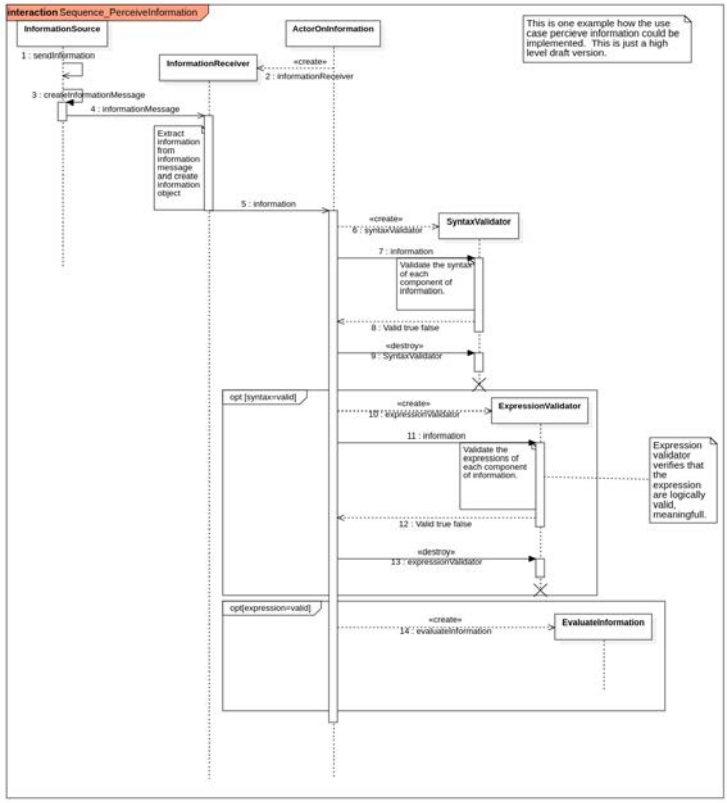

Figure 24: Interaction diagram of perceive information.

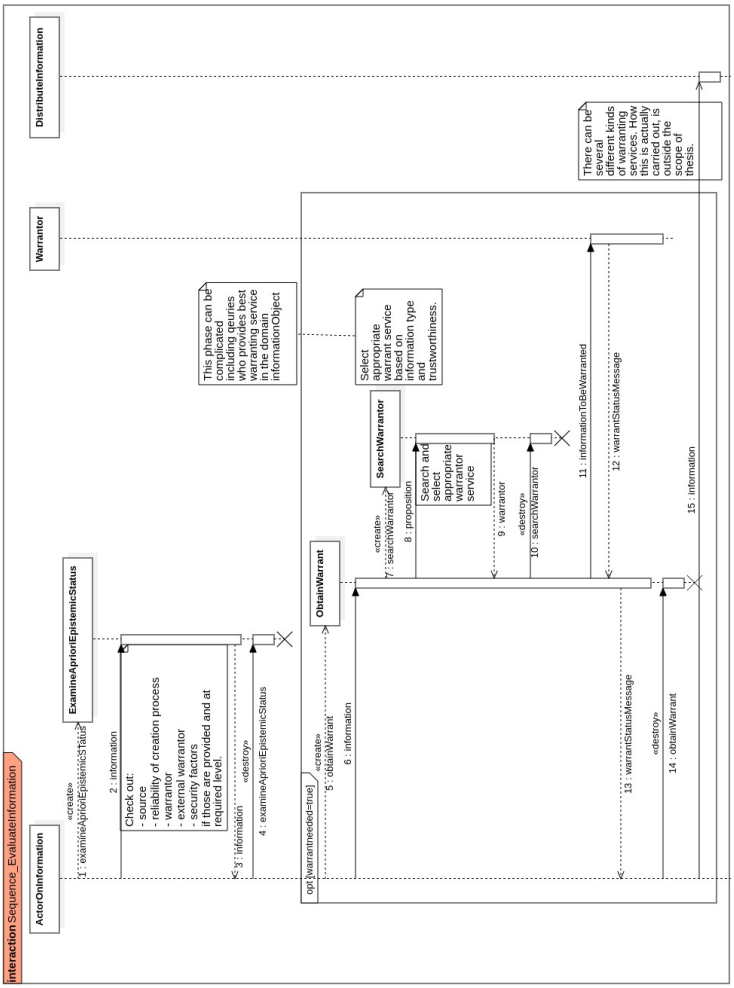

Figure 25: Interaction diagram of evaluate information.

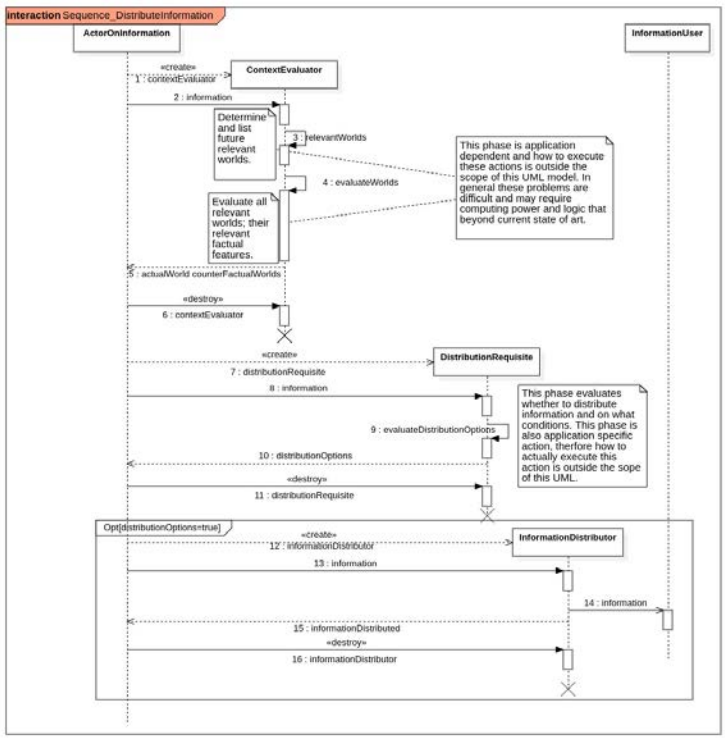

Figure 26: Interaction diagram of distribute information.

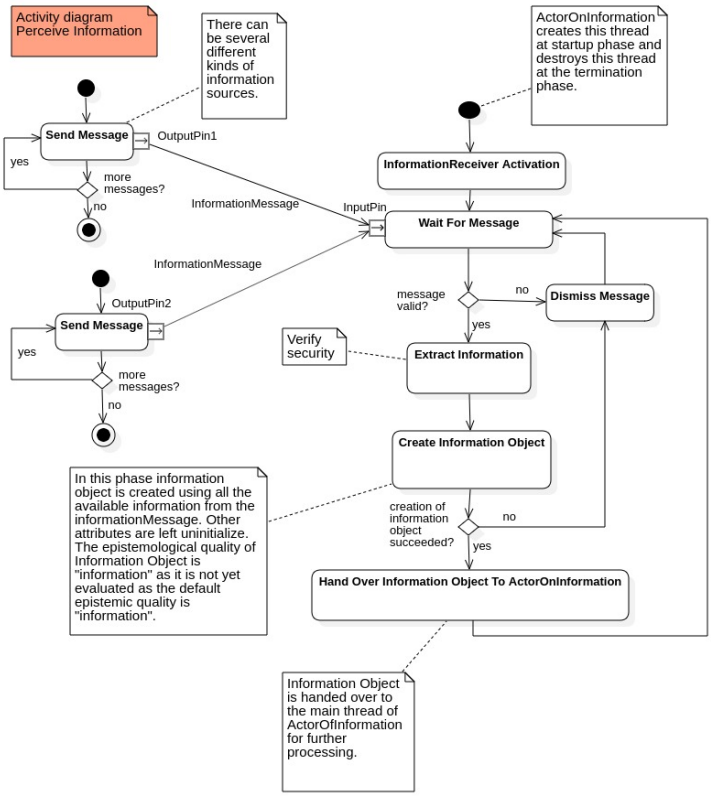

Figure 27: Activity diagram of perceive information.

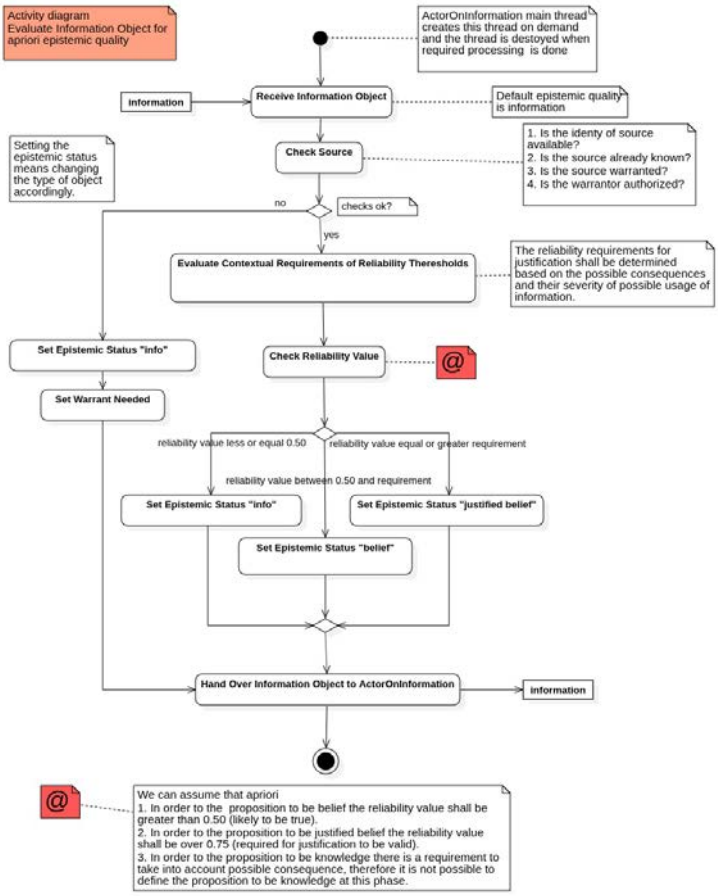

Figure 28: Activity diagram of evaluate information.

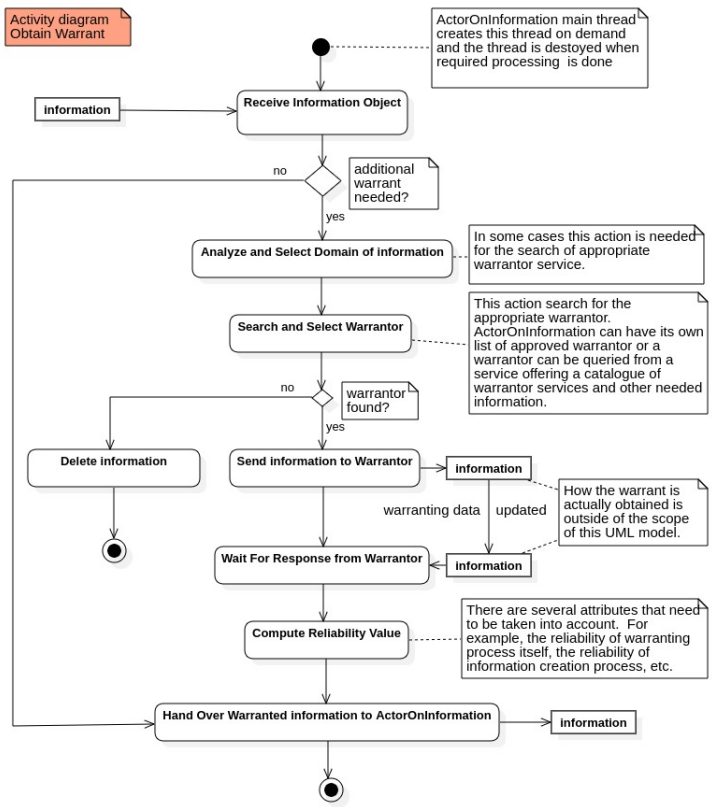

Figure 29: Activity diagram of obtain warrant.

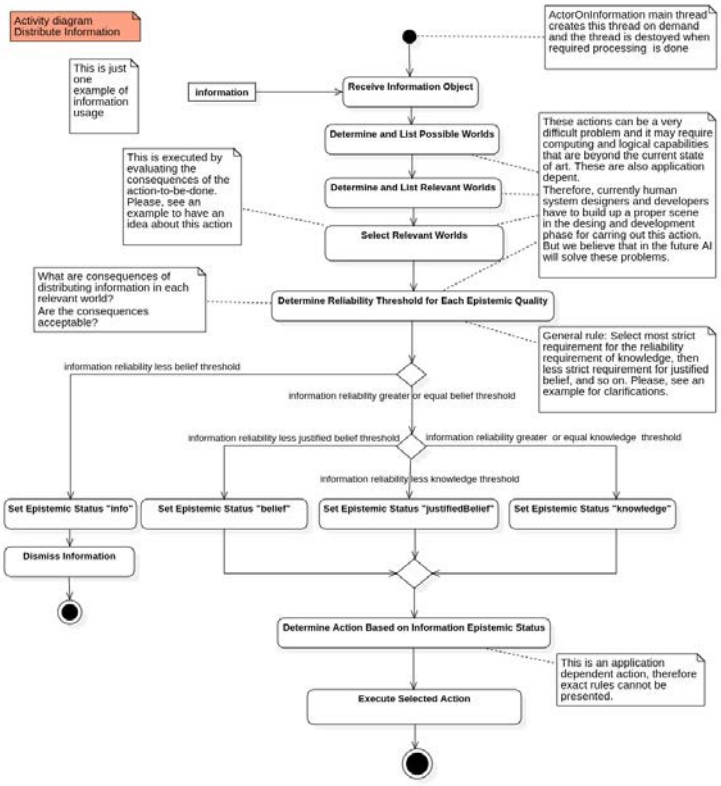

Figure 30: Activity diagram of distribute information.

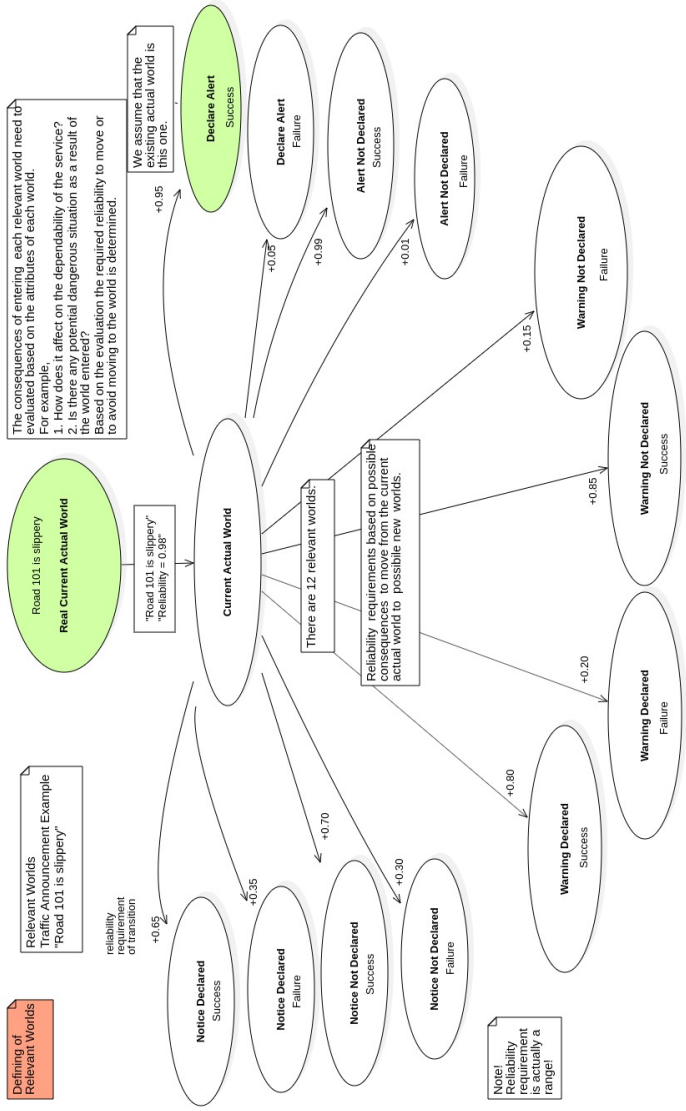

Figure 31: Example of specifying possible worlds and requirements set by possible worlds.

APPENDIX 3

—

DISCUSSIONS ON EVALUATING EPISTEMIC QUALITY OF BELIEFS

CONTENTS

1 Introduction 1
2 Defining Reliability^{p} Requirements 3
3 Sources of Beliefs^{p} 6
3.1 Sensor . . . 8
3.2 Inference Service . . . 10
3.3 Memory^{c} . . . 14
3.4 Distributed Information Base . . . 16
3.5 Warrant Service . . . 18
3.6 Human being . . . 22
3.7 Social Media (human beings) . . . 25
3.8 Common Information Service . . . 28
3.9 Another ISA_{bdi}-X . . . 30
4 Ontology of Traffic Information System 34
5 Summary 36
Appendices 38
A Ontology 38

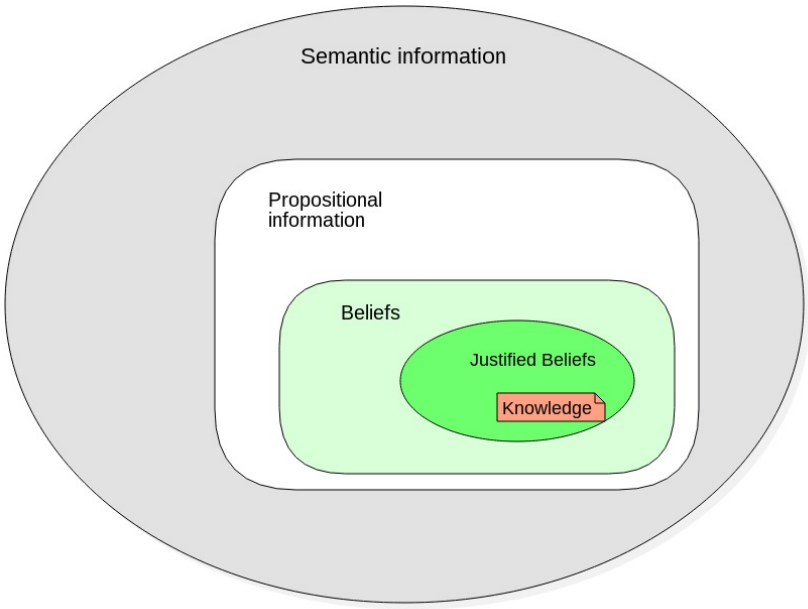

Figure 1: Different Classes of Beliefs.

1 INTRODUCTION

The epistemic quality of information has a significant role in human beings' actions and thoughts. We claim that in the case of intelligent software agents the quality of perceived information also has a very important role in the execution of actions whether the actions be a distribution of beliefs or physical activity. In this appendix we discuss specifying reliability^{p}[1] requirements for the epistemic quality of information, various sources of information, and the affect of the sources upon the epistemic quality of information perceived by intelligent software agents. We have defined that the epistemic quality of information depends on the reliability^{p} of belief–forming processes (based on [1, 2]) (and/or of warrant services) and the expected consequences of the utilization of belief. Different classes of information are illustrated in Figure 1. In order to illustrate different cases we use a sce-

1 There are several terms, which have different explication in philosophy and computer science. We use superscript 'p' when we use a term in philosophical meaning and superscript 'c' in the meaning of computer science.

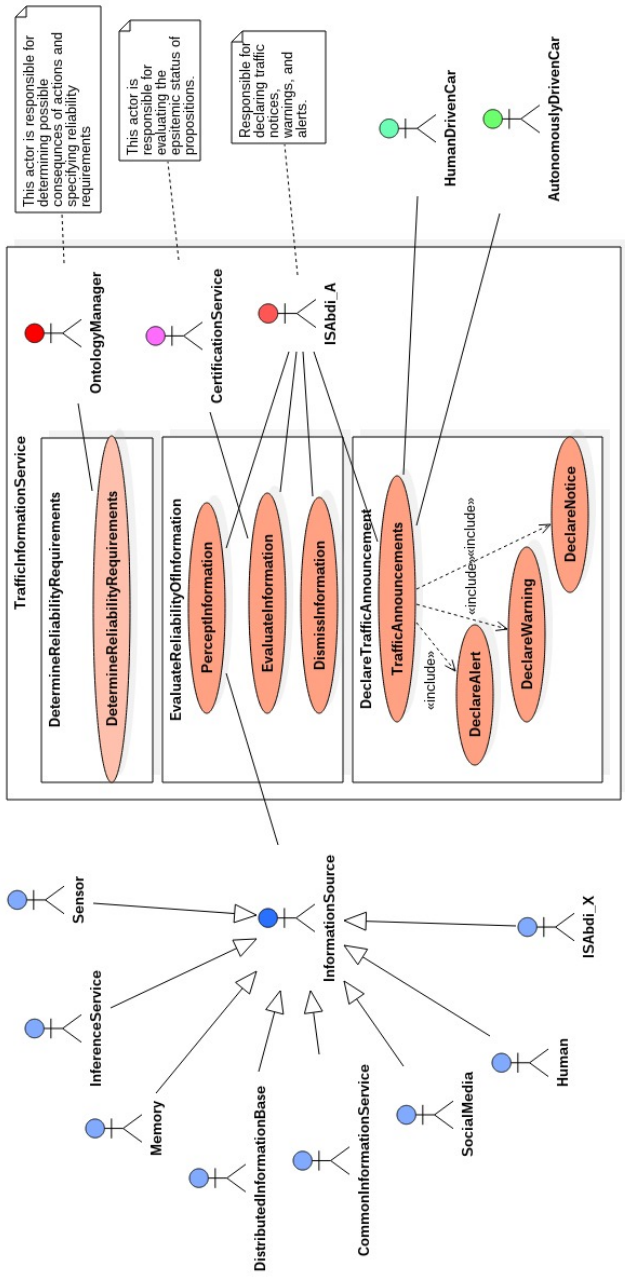

Figure 2: Use Case of Traffic Information Service.

nario of a traffic information service (hereinafter TIS[2]) as illustrated in Figure 2 in the form of UML use case diagram. TIS is provided by several co-operating intelligent software agents (hereinafter ISA$_{bdi}$s)[3] and human beings. The role of ISA$_{bdi}$-A is to announce traffic notices, warnings or alerts both to human drivers (hereinafter HDC) and autonomous cars driven by ISAs (hereinafter ADC), when cars are approaching Road 101 planning to enter into it, and the belief *"Road 101 is slippery."* fulfils specified epistemic requirements.

The scenario is divided into three sub–scenarios, which describe the determination of reliability^{p} values for belief, justified belief, and knowledge, the evaluating of information from different sources, and the declaration of a traffic announcement.

2 DEFINING RELIABILITY^{p} REQUIREMENTS

ISA$_{bdi}$-A should declare a traffic notice, when there is a belief (there is no acceptable justification available) that the road might be slippery. ISA$_{bdi}$-A should declare a traffic warning, when there is a justified belief that the road could be slippery, and ISA$_{bdi}$-A should declare a traffic alert, when there is knowledge that the road is slippery.

What are the reliability^{p} requirements of the belief creation processes for information to be belief, justified belief or knowledge? In order to answer this question we need, at first, to analyse the possible consequences of various traffic declarations. In general, in order to analyse possible consequences of an action requires that all the relevant possible worlds, which could be results of the action, need to be determined. As such, this is a very difficult problem—the so–called frame problem—comprising two different challenges: first, to determine the possible worlds may require resources that are beyond the power of contemporary logic as well as beyond current computing power; second, to determine the relevant worlds requires the analysis of relevance in the context of an application, and then the analysis of each possible world

2 This is purely a hypothetical example in order to clarify our thinking about the role of beliefs in this kind of environment.

3 We use ISA$_{bdi}$ subscript when we refer to the belief–intention–desire type of intelligent software agent.

♯	**Action**	**Reality - road is**
1	Declare traffic notice	slippery
2	Declare traffic notice	not slippery
3	Do not declare traffic notice	slippery
4	Do not declare traffic notice	not slippery
5	Declare traffic warning	slippery
6	Declare traffic warning	not slippery
7	Do not declare traffic warning	slippery
8	Do not declare traffic warning	not slippery
9	Declare traffic alert	slippery
10	Declare traffic alert	not slippery
11	Do not declare traffic alert	slippery
12	Do not declare traffic alert	not slippery

Table 1: Possible actual and counter–factual worlds of traffic information service.

regarding the relevance. Therefore, in most application domains these problems are still human resolvable ones. However, we believe that the research in artificial intelligence and in other related domains will resolve these problems in the future.

In our scenario there are 12 relevant worlds as listed in Table 1. The consequences of the actions, the outcomes of which are successful (# 1, 4, 5, 8, 9, and 12) are the ones that are expected from the system; therefore, we consider that in these cases the system is dependable. The other six worlds (# 2, 3, 6, 7, 10, and 11) require a more thorough analysis. Declaring a non-valid traffic notice, warning, or alert (worlds 2, 6, and 10) results in the consequence that the dependability of the system decreases, and trust in the system suffers. But not declaring a traffic notice or warning when it should done may result in dangerous situations in traffic. And not declaring a traffic alert when it should be done (world # 11) can result in situations, where traffic accidents are likely to happen. Therefore, the world # 11 sets the strictest requirements for the epistemic quality of beliefs. We assume that human beings as drivers are more capable to adapt themselves to

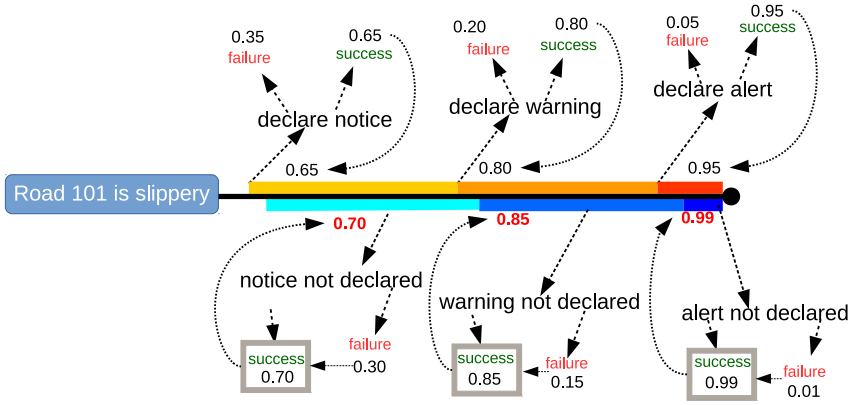

Figure 3: Evaluation of Consequences (ADC).

road conditions than ISAs as drivers. Hence, the requirements for the epistemic quality of beliefs are more rigorous in the case of ADC. We also assume that ADCs are guided not to enter a road, where a traffic alert has been declared. As TIS has to produce epistemically high quality beliefs (including the metadata expressing the reliability^p of the process producing the beliefs), it shall operate on such formed beliefs, with which it is able to produce the required epistemic quality.

Let us assume that the requirements illustrated in Figure 3 and detailed in Table 2 are set to TIS.

As mentioned above failing to declare a notice, warning, or alert might result the most severe consequences, therefore, these set the highest requirements for the epistemic quality, that is the highest requirements for the reliability^p of the processes producing beliefs. The worlds # 4, 8, and 12 set the highest requirement for belief, justified belief and knowledge and therefore, for the reliability^p of the processes of TIS. We can consider that in order for the piece of information stated by proposition *"Road 101 is*

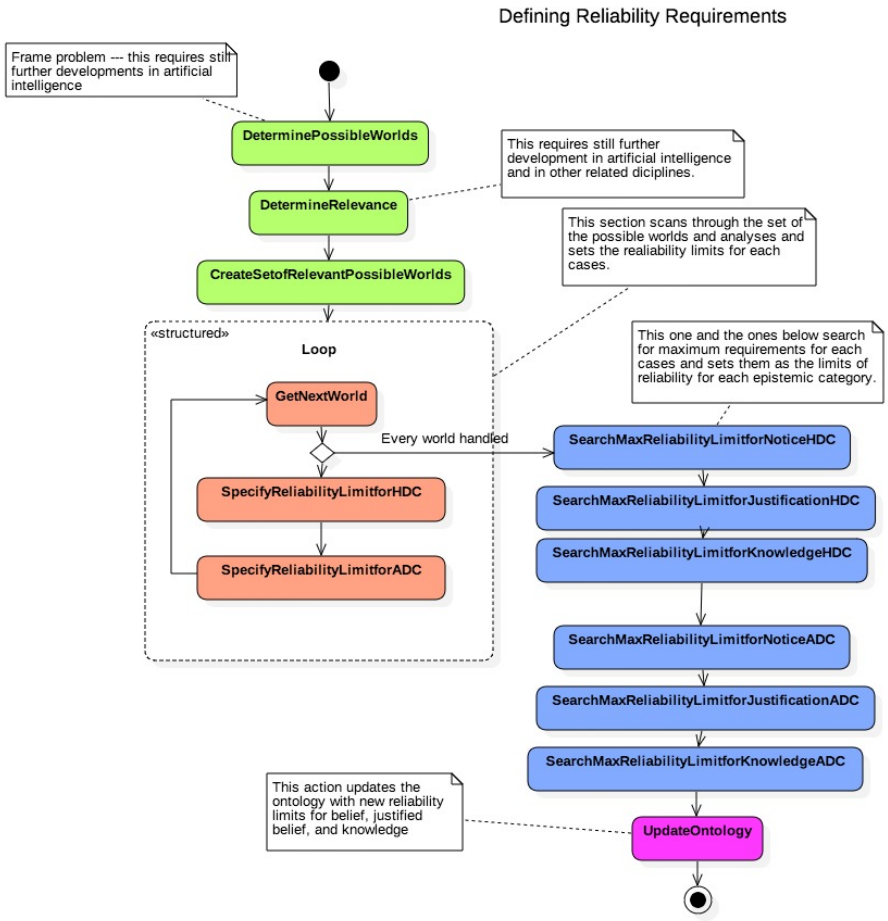

Figure 4: Specifying Reliability Requirements.

slippery." to be the object of a belief, justified belief, or knowledge demands the reliability^{p} of the processes of producing information to be 0.70, 0.85, or 0.99 respectively.

We can summarize the above discussion in the form of UML process description as illustrated in Figure 4.

3 SOURCES OF BELIEFS^{p}

There are several different kinds of information sources, and next we discuss some of their specific features related to the epistemic

#	Action	Status	ADC – Reliability^{p}	HDC – Reliability^{p}
1	Declare notice	success	0.65	0.60
2	Declare notice	failure	0.35	0.40
3	Don't declare notice	success	**0.70**	**0.65**
4	Don't declare notice	failure	0.30	0.35
5	Declare warning	success	0.80	0.74
6	Declare warning	failure	0.20	0.26
7	Don't declare warning	success	**0.85**	**0.77**
8	Don't declare warning	failure	0.15	0.23
9	Declare alert	success	0.95	0.90
10	Declare alert	failure	0.05	0.10
11	Don't declare alert	success	**0.99**	**0.95**
12	Don't declare alert	failure	0.01	0.05

Table 2: Reliability^{p} requirements for declarations.

quality of the piece of information stated by proposition *"Road 101 is slippery."*. Figure 5 illustrates the possible main sources of information, from which ISA$_{bdi}$-A may perceive information either through data communication services or directly using different input mechanisms, such as internal application interface, keyboard interface, mouse interface, video camera interface, sensor interface, etc.

Let us assume, first, that Road 101 is in fact slippery (world #9 in Table 1 is the actual world; thus, all other ones are counter–factual worlds), and ISA$_{bdi}$-A should declare a traffic alert,[4] and second, that ISA$_{bdi}$-A either perceives the piece of information stated by the proposition *"Road 101 is slippery."* or required information to infer it from the sources illustrated in Figure 5.

3.1 Sensor

In this case (illustrated in Figure 6) we assume that on Road 101 there is an intelligent road sensor system that can detect the slipperiness of the road. The manufacturer of the road sensor system has obtained from a certification institute a warrant that the reliability^{p} of the sensor system is 0.99. ISA$_{bdi}$-A perceives from the road sensor "Road 101 is slippery." with associated metadata *"Reliability is 0.99." "Warranted by VTT"*. In our example this does fulfil the reliability^{p} requirement for the formed belief to be knowledge in both cases (ADC and HDC). Thus, ISA$_{bdi}$-A declares the traffic alert both to ADCs and to HDCs.

Conclusion: In this case we consider ISA$_{bdi}$-A to be dependable.

The key issues are as follows:

1. Theory: reliabilism
2. Reliability can be checked from a certification institute or in some cases from a manufacturer (if it is considered to be trustworthy, enough).

4 We consider this to be the actual circumstance; thus the proposition "Road 101 is slippery" is true.

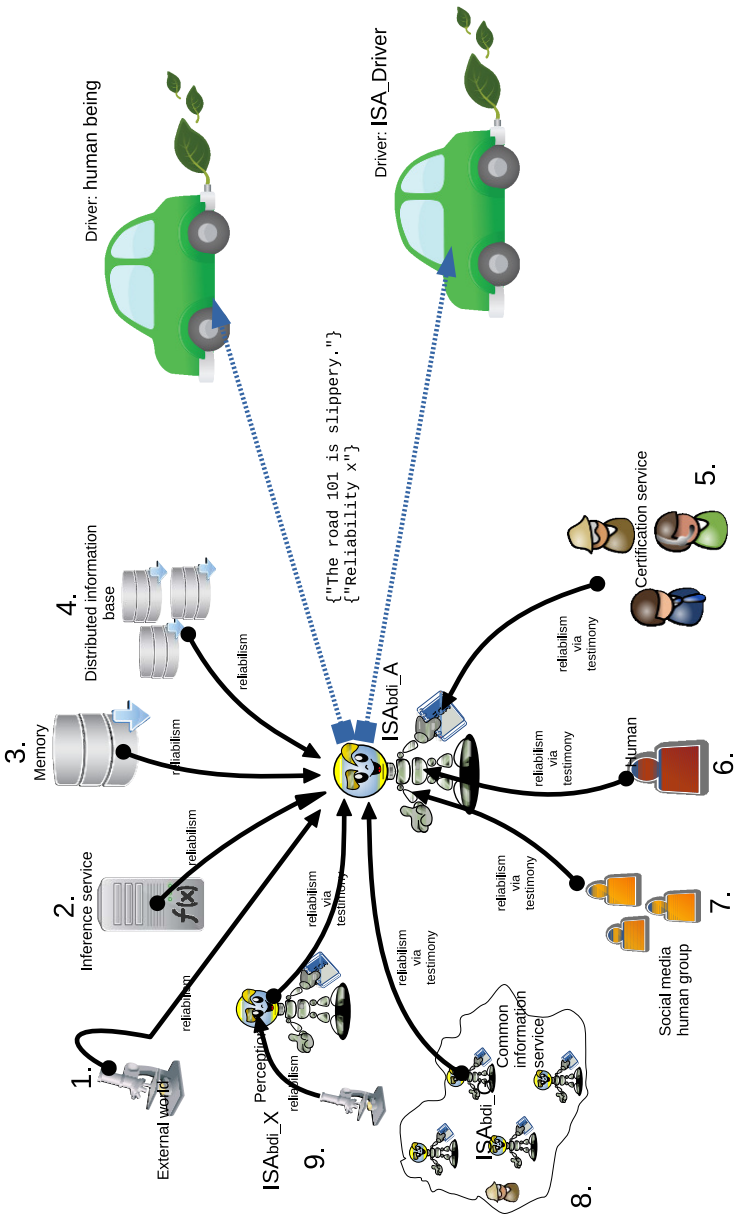

Figure 5: Sources of Information of ISA.

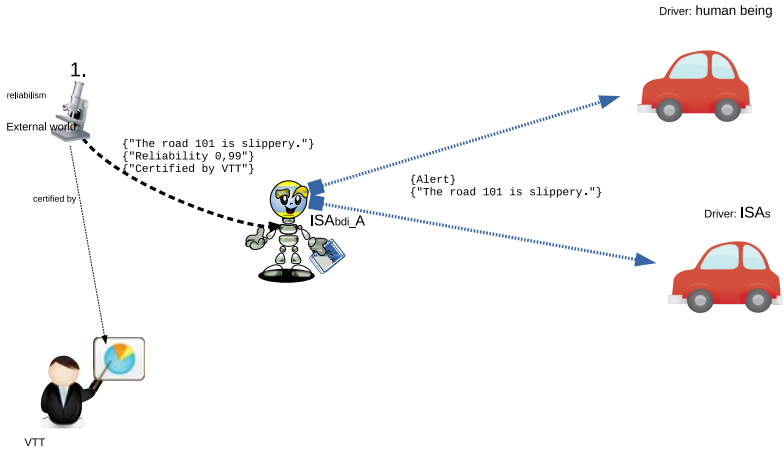

Figure 6: Sensor as the Source.

3.2 Inference Service

In this case (illustrated in Figure 7) we assume that ISA$_{bdi}$-A perceives information from six different sensors and utilizes an inference service to infer whether or not to declare a traffic warning. Let us further assume that the reliability^{p} of the sensors have been tested and warranted, and the results are available to ISA$_{bdi}$-A.

There are several different kinds of inference engines starting from simple "IF-THEN" rule engines ending at either more powerful theorem provers implementing various modal logics or neural networks. Each of these may have different levels of reliability^{p} to produce beliefs. The reliability^{p} of the inference service is used to determine the epistemic quality of the result of an inference process, where affecting factors are, for instance, the reliability^{p} of inference algorithms and propositions used.[5] Other factors to determine reliability^{p} can be derived from the dependability theory of computer science and the logics used by inference engines (for example truth–preserving logics).

5 How this is actually done is an application dependent factor and thus it is outside of the scope of this thesis.

Let us assume that ISA$_{bdi}$-A perceives from the sensors the following data and requests the inference service to analyse the condition of Road 101:[6].

1. *"Road 101." "Reliability 1.0" "Warranted by Trafi"*
2. *"Air temperature is -3°C." "Reliability 0.99" "Warranted by VTT"*
3. *"Road temperature is -1°C." "Reliability 0.99" "Warranted by VTT"*
4. *"It is snowing." "Reliability 0.99" "Warranted by VTT"*
5. *"The depth of snow is 2cm." "Reliability 0.95" "Warranted by VTT"*
6. *"The dew point is -3°C." "Reliability 0.99" "Warranted by VTT"*
7. *"The road is salted." "Reliability 1.0" "Warranted by Trafi"*.

Let us also assume that the reliability^{p} of the inferring process itself is 0.9999, which is warranted by the developer of the inference service.

All other factors fulfil the reliability^{p} requirement for knowledge, except the fifth one, which plays a specific role, here. It can be considered to be either knowledge or justified belief. If the level of the reliability^{p} requirement is transitive (the reliability^{p} 0.99 is required also from all input beliefs) then the proposition *"The depth of snow is 2cm."* is considered to be the object of justified belief but not knowledge. Thus, it is the decisive factor, and the inference system infers that the reliability^{p} of the result of the reasoning process in this kind of environment is 0.98. It responds to ISA$_{bdi}$-A with the following belief provided with a metadata:

"Road 101 is slippery." "Reliability is 0.98."[7]

In the case of ADC this does not fulfil the reliability^{p} requirement for the piece of information stated by the proposition *"Road 101 is slippery."* to be knowledge. Therefore, ISA$_{bdi}$-A does not

6 We assume also that the reliability^{p} (the accuracy of the equipment) values are provided by sensor manufacturers.

7 Note that in this case the reliability^{p} is not inferred directly from the probability calculation.

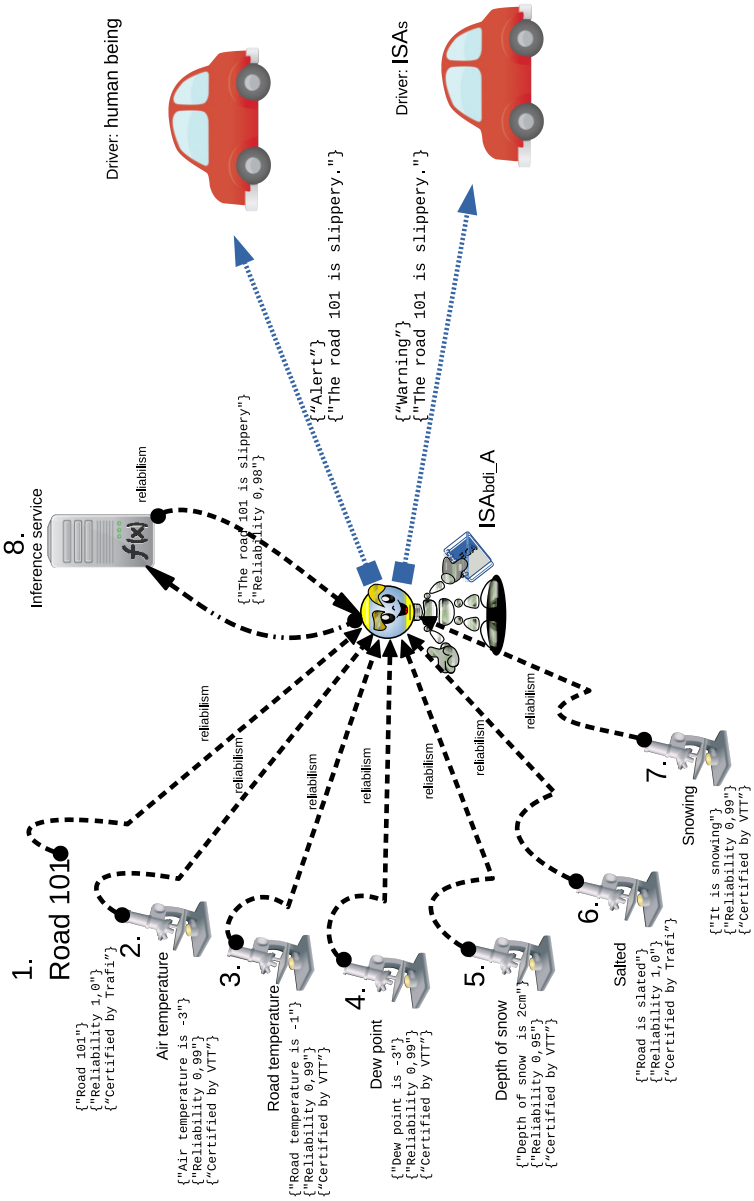

Figure 7: Sensors and Inference Service.

declare a traffic alert but a traffic warning, even though the traffic alert should be declared. However, in the case of HDC the reliability of 0.98 fulfils the requirement for the piece of information stated by the proposition *"Road 101 is slippery." "Road 101 is slippery."* to be knowledge, and ISA$_{bdi}$-A declares a traffic alert, which is valid.

Conclusion: In this case our intuition says that ISA$_{bdi}$-A is dependable in the case of HDC, but not in the case of ADC. But ISA$_{bdi}$-A operates according to its specifications. This contradiction clearly indicates that the epistemic quality of information affects the dependability.

1. *"Road 101." "Reliability 1.0" "Warranted by Trafi" "Knowledge requirement limit 1.0"*
2. *"Temperature is -1°C." "Reliability 0.99" "Warranted by VTT" "Knowledge requirement limit 0.99"*
3. *"It is snowing." "Reliability 0.99" "Warranted by VTT" "Knowledge requirement limit 0.99"*
4. *"The depth of snow is 2cm." "Reliability 0.87" "Warranted by VTT" "Knowledge requirement limit 0.80"*
5. *"The dew point is -3°C." "Reliability 0.99" "Warranted by VTT" "Knowledge requirement limit 0.99"*
6. *"The road is salted." "Reliability 1.0" "Warranted by Trafi" "Knowledge requirement limit 1.0"*.

If the reliability^{p} requirement is not transitive, and each factor has its own reliability^{p} limit based on the type of the sensor, then the outcome changes. Now, each belief has its own weight value (the belief *"The depth of snow is 2cm."* has the lowest value) in the inference process. Now, the inference system infers that the reliability^{p} of the result of the reasoning process in this kind of environment is 0.99. This does fulfil the reliability^{p} requirement for the piece of information stated by the proposition *"Road 101 is slippery."* to be knowledge in both cases. Therefore, ISA$_{bdi}$-A declares the traffic alert both to HDC and to ADC.

Conclusion: In this case we consider ISA$_{bdi}$-A to be dependable.

The key issues are as follows:

1. Theory: reliabilism.
2. The same reliability^{p} of process can produce in one instance knowledge and in another one only justified belief.
3. It is possible that the epistemic quality of the perceived piece of information can create a situation, where a system is dependable in one instance and not dependable in another instance.
4. The evaluation of the possible consequences can be a difficult task in reality—comprises the frame problem and requires the management of counter–factual worlds.
5. How do we take into account the effect of the reliability^{p} of various processes, when evaluating the overall reliability^{p} of the process?

3.3 Memory^{c}

In this case (illustrated in Figure 8) we consider the memory^{c} to be ISA$_{bdi}$-A's private one, which is accessible only to either the developers/maintainers of ISA$_{bdi}$-A or ISA$_{bdi}$-A itself. There are two ways to store beliefs^{p} into the memory^{c}: either a developer/maintainer stores them or ISA$_{bdi}$-A itself stores them. In both cases the reliability^{p} needs to be evaluated and stored in the metadata of the belief. The case of ISA$_{bdi}$-A itself storing the belief is straightforward, because ISA$_{bdi}$-A knows the origin and the base of the justification (e.g. an inference service). The case of the developer/maintainer is more complicated, and it demands that the developer/maintainer of belief specifies and collects the required justification of the belief.

Let us assume that ISA$_{bdi}$-A has stored for later utilization the following belief: *"Road 101 is slippery."* with associated metadata *"Reliability 0.99" "Time 02.03.2020 14:00" "Source ISA$_{bdi}$-A"*. Let us further assume that ISA$_{bdi}$-A retrieves at 02.03.2020 14:01 the

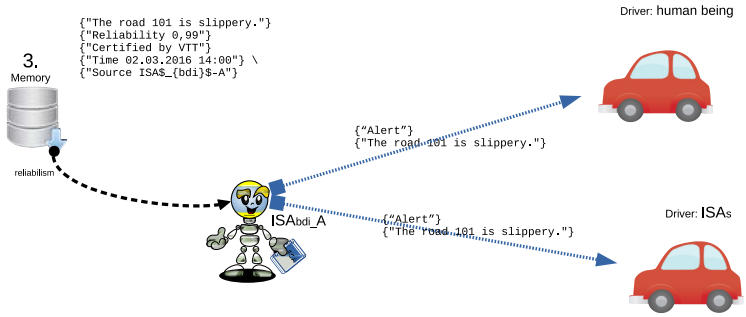

Figure 8: Memory.

belief stated by the proposition *"Road 101 is slippery."* with the associated metadata, and ISA$_{bdi}$-A declares the traffic alert.

But shall ISA$_{bdi}$-A declare the traffic alert? In this case there is a time factor that needs to be taken into account in the evaluation of the epistemic quality of the belief.[8] There are two kinds of beliefs: volatile and constant.[9]. A constant belief maintains its reliability^{p} value independent of time. As an example of this kind of a belief we have the following one: *"2 + 2 = 4" "Reliability is 1.0."* The belief *"Road 101 is slippery."* is volatile, because its epistemic quality may change depending on the time of utilizing the belief. The belief may gain or lose its justification as time is passing. In this case our intuition says that a one minute delay does not change the justification.

8 There are other factors, such as the reliability^{p} of the memory itself and the data transfer bus of a computer, which could be taken into account, but we consider that their reliability^{p} in this context is so high that they do not affect the result in any way.

9 The related logical terms are contingent and tautology.

Conclusion: In this case we consider ISA$_{bdi}$-A to be dependable.

Let us assume that a system manager has stored the following belief: *"Road 101 is slippery."* with associated metadata *"Reliability 0.99" "Time 02.03.2020 14:00" "Source Manager-A"*.

ISA$_{bdi}$-A retrieves at 02.03.2020 14:01 from the memory the belief *"Road 101 is slippery."* with the associated metadata, and ISA$_{bdi}$-A declares the traffic alert.

But should ISA$_{bdi}$-A announce the traffic alert in this case? This is an example of testimonially transferred—via a memory—belief. In this case our intuition says that this is a special case of testimony, because Manager-A has a special role in the context of ISA$_{bdi}$-A. Manager-A is a kind of warrant source of information in its relation to ISA$_{bdi}$-A. Therefore, our intuition says that ISA$_{bdi}$-A does not need any other justification for the belief to be knowledge.

Conclusion: In this case we consider ISA$_{bdi}$-A to be dependable.

The key issues are as follows:

1. Theory: reliabilism and in some cases reliabilism via testimony.
2. The time factor needs to be taken into account: stable and changeable epistemic values.
3. The role of developers and system managers as 'warranted' sources of beliefs. Do they know the reliability^{p} of their own processes?

3.4 Distributed Information Base

In this case (illustrated in Figure 9) we assume that ISA$_{bdi}$-A has stored for later utilization the belief *"Road 101 is slippery."* with required metadata (reliability factor, time stamp, etc.) into the distributed information base: *"Road 101 is slippery."* with associated metadata *"Reliability 0.99" "Time 02.03.2020 14:00" "Source ISA$_{bdi}$-A"*.

At least the following factors need to be considered when determining the epistemic quality of the proposition:

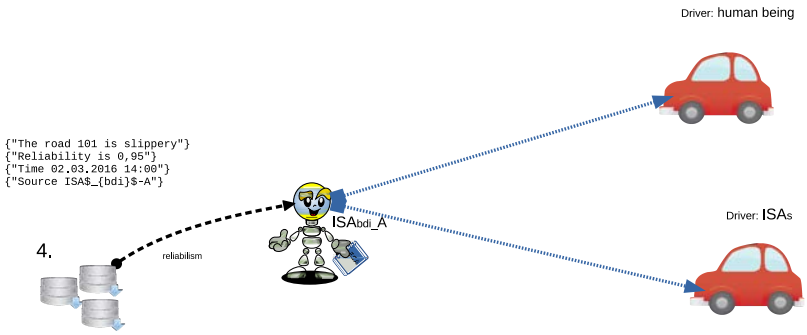

Figure 9: Distributed Information Base.

1. The reliability^{p} of processes that were used to established the belief.
2. The reliability^{p} of processes that are used to store the belief (the process to populate the information database).
3. The reliability^{p} of information storing (integrity).
4. The coherence of the different information sets.
5. The up-to-date status of information.
6. The role of testimony.

How can these values be obtained? The first case is a straightforward one: as discussed in examples 1 and 2 above the reliability^{p} is 0.99. In cases 2 – 4 the reliability^{p} could be obtained from the specification of the distributed information base.[10] The fifth case deals with the time factor as discussed in case 3 above.

10 This demands that the developers, the supervisors, or a certification institute specifies and collects the required data.

Let us suppose that ISA$_{bdi}$-A retrieves at 02.03.2020 20:00 the belief *"Road 101 is slippery."* with associated metadata *"Reliability 0.99" "Time 02.03.2020 14:00" "Source ISA$_{bdi}$-A"* ISA$_{bdi}$-A does not declare a traffic alert. But should ISA$_{bdi}$-A declare the traffic alert? Due to the changing nature of weather, the traffic alert system must operate as a real time system; therefore, our intuition says that there is no longer justification for the belief *"Road 101 is slippery."*. ISA$_{bdi}$-A shall not declare a traffic alert, warning, or notice based on this belief.

Conclusion: In this case we consider ISA$_{bdi}$-A to be dependable.

The key issues are as follows:

1. Theory: reliabilism and in some cases reliabilism via testimony.
2. There are other factors related to belief in addition to the reliability^{p} of the process that affect the dependability of the system, such as the time factor, which needs to be taken into account.
3. The question about the role of testimony: Many various sources may store beliefs into the distributed information base; therefore, the source of the belief needs to be stored into the distributed information base with associated metadata expressing the reliability^{p} of the belief–forming process.
4. Testimony is not the source of justification or knowledge, but only a transmission method.
5. The evaluation of the reliability^{p} of the processes of distributed information bases can be difficult; for example, commercial cloud services do not provide such data.

3.5 Warrant Service

In this case (illustrated in Figure 10) we assume that ISA$_{bdi}$-A perceives from a source X the piece of information stated by the

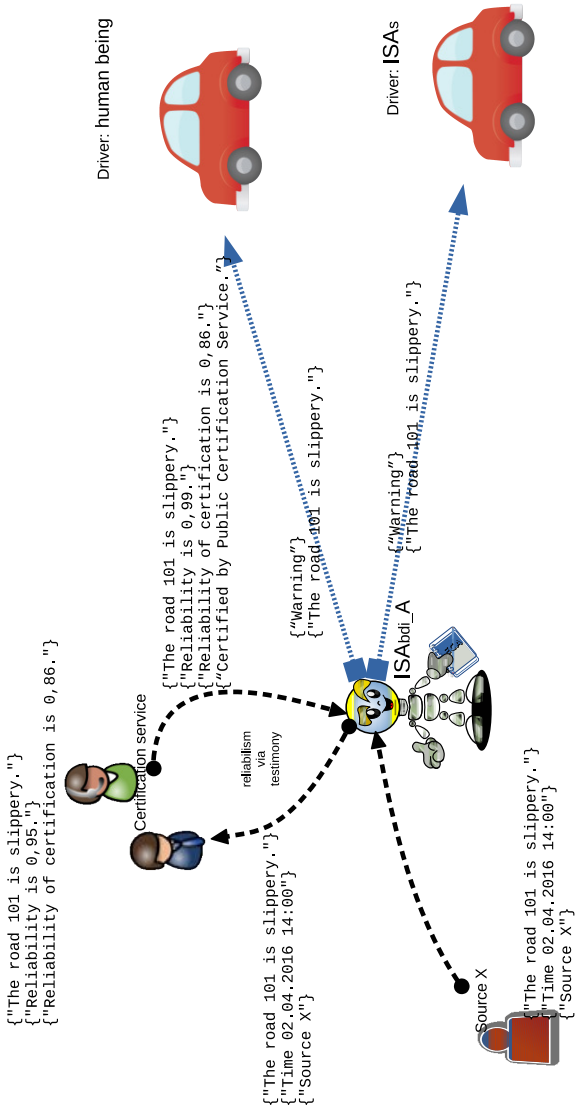

Figure 10: Warrant Service.

proposition *"Road 101 is slippery."* with associated metadata *"Time 02.04.2020 14:00" "Source X"*. Because there is no reliability^{p} data available, ISA_{bdi}-A sends the belief to a warrant service to be warranted.

Warrant service is an important business area dealing with trust; if something is warranted, there is an indication that it can be trustworthy. We consider a warrant to provide a justification for belief. There are several different kinds of warrant services providing warrants in numerous domains, such as healthcare (e.g. the Joint Commission), education (e.g. universities), IT profession (e.g. Microsoft), and legal profession (e.g. universities). In the domain of distributed computing systems digital certificate services (utilized in cryptography: the certification authority certifies the identity of the peer in a trusted relationship) are the most common ones having an administration infrastructure including public governing bodies.[11]

In this case we refer to the domain of information warrant service (hereinafter IWS), that evaluates the epistemic quality of information based on various factors including the reliability^{p} of the process, which has produced information. In the Web there are several similar kinds of services, which are called "fact checking" services[12] [3]:

- Full Fact (www.fullfact.org) which is an independent, non–partisan, fact–checking charity operating in UK
- FactCheck.org (www.factcheck.org) which is a project of the Annenberg Public Policy Center
- Fact Checker (www.washingtonpost.com/news/fact-checker) which the Washington Post offers to its readers
- PolitiFact (www.politifact.com) which is a project of the Tampa Bay Times
- Snopes.com (www.snopes.com) which is founded by David Mikkelson
- TruthOrFiction (www.truthorfiction.com) which is provided by Branches Communications, Inc. USA.

11 EU commission has listed authorized digital certificate service providers.

12 This is the case especially in the United States.

These "fact checking" services operate mainly in the domains of politics and public press, and facts are usually checked manually by human beings.[13] However, we argue that in the future "fact checking" services will be expanded into other domains, which demand high epistemic quality of information. And the evaluation will also be carried out by autonomous, intelligent warrant agents. But, so far it is a real challenge to create a fully automated information quality evaluation system [4], and this topic is outside of the topic of this thesis.

The main objective of using the warrant service is to obtain justification for the belief in the form of the epistemic quality of the piece of information stated by a proposition. What is required from the warrant service? Firstly, the warrant service must evaluate the reliability^{p} of the process that has created information. However, there are instances, in which the evaluation cannot be done reliably^{p} enough, because required information is not available. Secondly, the warrant service needs to know the reliability^{p} of its own warranting process.

Let us assume that after the evaluation ISA$_{bdi}$-A perceives from the warrant service: *"Road 101 is slippery." "Reliability^{p} is 0.95." "Reliability^{p} of warrant is 0.86." Warranted by Public Warrant Service."*. The third one expresses the reliability^{p} of the process of establishing the warrant. Now, there are two separate factors to be taken into account when inferring whether or not to announce a traffic warning. In our scenario this does not fulfil the reliability^{p} requirement for the belief to be knowledge, as the reliability^{p} of the evaluation process is not high, enough. But it is high enough for the belief *"Road 101 is slippery."* to be justified belief. ISA$_{bdi}$-A declares the traffic warning both to ADC and to HDC.

But is ISA$_{bdi}$-A dependable? The question is raised, because in the actual world the traffic alert should be declared. ISA$_{bdi}$-A operates on the belief, which the reliabilities^{p} of the belief forming and warrant processes are not adequate for it to be knowledge. And as specified in the scenario ISA$_{bdi}$-A shall declare the traffic warning, when the belief fulfils the requirement^{p} of justified belief but does not fulfil the requirement^{p} of knowledge. Therefore, ISA^{p}-A operates according to its specifications, but due to the lack of epistemic quality of the perceived piece of information

13 This is the case at at time of writing this thesis.

ISA$_{bdi}$-A is not dependable from the viewpoint of users.

Conclusion: In this case we consider ISA$_{bdi}$-A not to be dependable. But the main reason for not being dependable is the lack of epistemically high enough quality information. In this sense ISA$_{bdi}$-A is equivalent to human operated system: without epistemically proper beliefs it is difficult to act dependably.

The key issues are as follows:

1. Theory: reliabilism and reliabilism via testimony.
2. Who can provide warrant services?
3. There are two factors: the reliability^{p} of the belief forming process, and the reliability^{p} of the process, which warrants the belief.
4. The evaluation of the reliability^{p} of the warrant process. An institute?
5. Is the warrant adequate in the cases, where the reliability^{p} of the belief forming process is not available, at all?
6. We can consider that warrant is the solution in the cases, where there are no direct data available about the reliability^{p} of the belief producing processes.

3.6 Human being

In this case (illustrated in Figure 11) we assume that ISA$_{bdi}$-A perceives the piece of information stated by the proposition *"Road 101 is slippery."* from a human being via a specific interface, such as a keyboard, voice, or short message service. There is a requirement to know the reliability^{p} of the person producing the piece of information. There are several factors, which can be used in the evaluation:

1. The accuracy of the person in past action
2. The appropriateness of the person to issue the piece of information

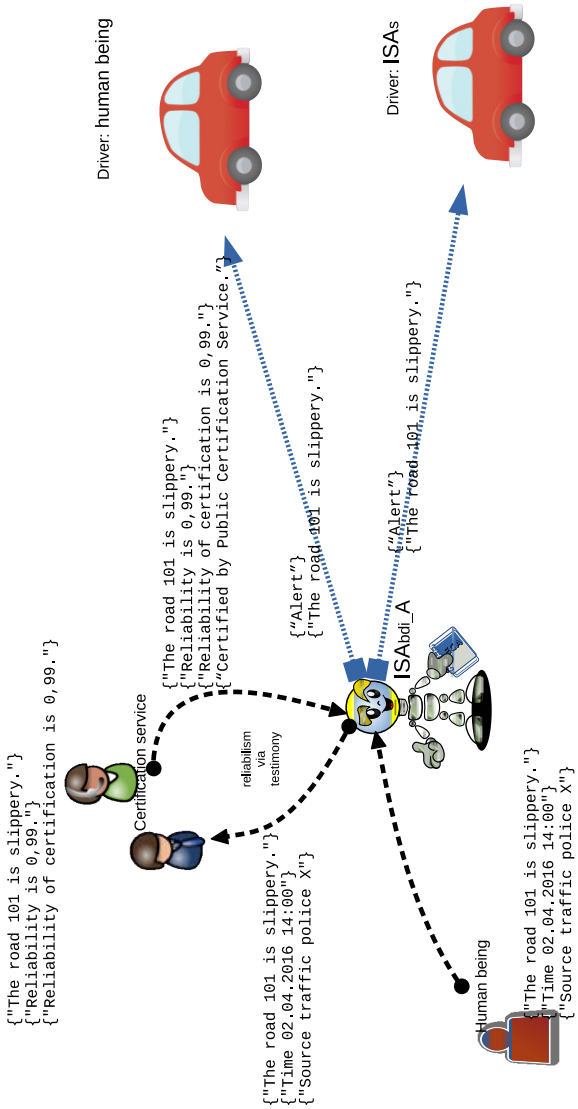

Figure 11: Human Being via warrant Service.

3. The consistency of the person in past actions
4. The relevance of the person to issue the piece of information
5. The reputation of the person (history)
6. The timeliness of the piece of information.

How the evaluation itself should actually be carried out is outside the topic of this discussion. There are some similarities with the evaluation of information quality, which has been discussed, for example, in [5, 6, 7, 8, 9, 10]. Another approach is the warrant of persons (for example, the reliability^{p} of the process of medical professors to form beliefs in the domains of their expertise.)

Let us assume that ISA$_{bdi}$-A perceives the piece of information stated by the proposition *"Road 101 is slippery."* from a traffic police X, and sends the following data to a warrant service for evaluation: *"Road 101 is slippery."* , *"Time 02.04.2020 14:00"*, *"Source traffic police X"*. Let us assume that the reliability^{p} of the traffic police to produce traffic information has been warranted to be 0.99. The traffic police X is a member of the traffic police, hence we consider the traffic police X's reliability^{p} to be 0.99, as well. After the evaluation ISA$_{bdi}$-A perceives from the warrant service: *"Road 101 is slippery." "Reliability^{p} is 0.99." "Reliability^{p} of warrant is 0.99." "Warranted by Public Warrant Service."*. Based on the result of the warrant service, ISA$_{bdi}$-A declares the traffic alert both to ADC and HDC.

Conclusion: In this case we consider ISA$_{bdi}$-A to be dependable.

The key issues are as follows:

1. Theory: reliabilism and reliabilism via testimony.
2. The reliability^{p} of a human being to produce belief.
3. The evaluation or warrant of a human being.
4. The evaluation of the reliability^{p} of the warrant process.
5. The warrant is the solution in the case of human beings (either beforehand or in real time).
6. The warrant of individual human being is a difficult issue. Professional individuals?

7. Does an individual inherit the reliability^{p} of the processes of an organization?

3.7 Social Media (human beings)

In this case (illustrated in Figure 12) we assume that ISA_{bdi}-A perceives the piece of information stated by the proposition *"Road 101 is slippery."* from a social media.

We refer social media to be a group of Internet (Web–based) applications that enable anyone to publish and access information, collaborate, and build relationships. There are several different kinds of social media, such as streams (Twitter), discussion boards (blogs) and forums (Google Groups), social networks (Facebook), reviews and ratings (Amazon.com), wiki (Wikipedia), wisdom of the crowd (Reddit), and questions and answers (Answers.com). Each of these has a different profile of the epistemic quality of information. For example, the profile of Wikipedia designates to justified beliefs and knowledge and the profile of Google Groups designates to beliefs (opinions).

Social media has achieved enormous popularity having over one billion users. This has resulted in a huge amount of user-generated information, because everyone can be an information producer. And the high majority of this information has been contributed by sources that do not/cannot provide any valid justification for beliefs. The epistemic quality of user–generated information can vary significantly from malicious false beliefs to highly credible justified beliefs (knowledge).

In the year 1997 Strong et.al. suggested that quality of information should be established during the manufacturing of the information [11], but this idea has not gained any popularity in social media. Social media applications do not provide users with any metadata about the reliability^{p} of the process that has created a published belief. Therefore to assess the epistemic quality of beliefs in social media is a huge problem, which is worth separate research projects. A similar problem is the trustworthiness of information in social media, which has been widely studied in various research projects and presented in a considerable number of articles, for example, in [5, 6, 7, 8, 9, 10, 12, 13].

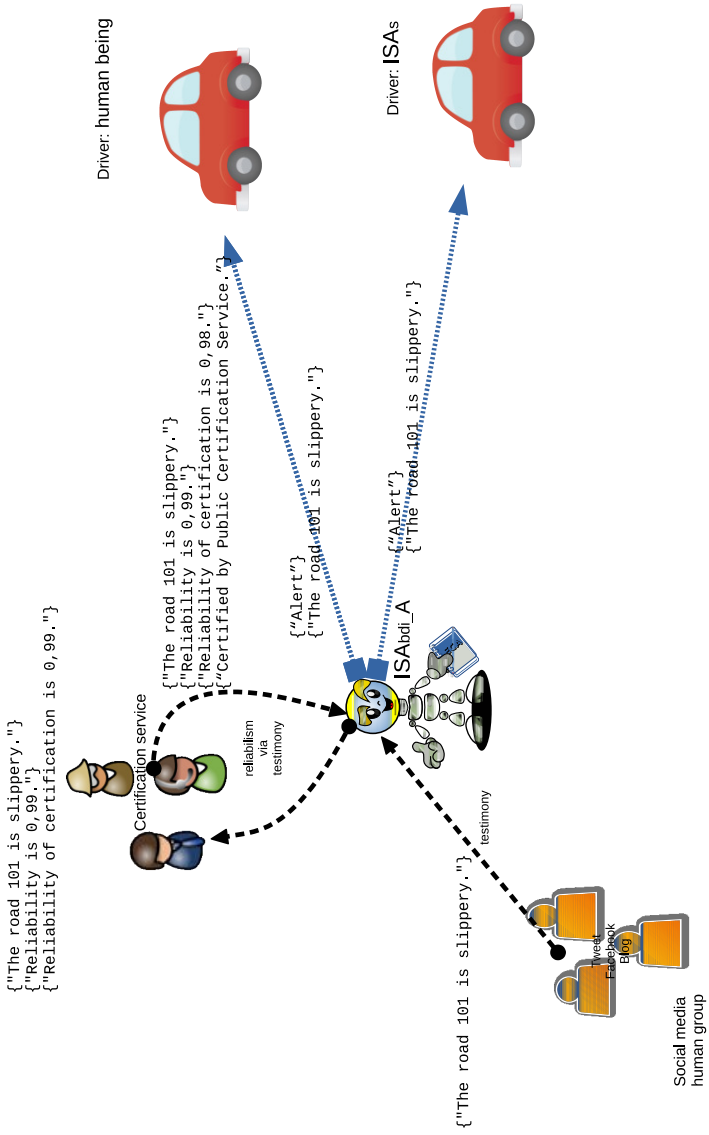

Figure 12: Social Media via Warrant Service.

We argue that testimony is not the source of either justification or knowledge; therefore, a priori epistemic quality of information received from social media is unknown. There is a possibility that the source of information does not himself/herself/itself believe the piece of information that he/she/it distributes. There is also a possibility that the piece of information has been maliciously altered. The user of the piece of information either needs to evaluate possible justifications, or requests a warrant service to carry out the evaluation.

Let us assume that ISA$_{bdi}$-A perceives the piece of information stated by the proposition *"Road 101 is slippery."*

1. as a tweet from a person X
2. as a notice from a Facebook group "101 drivers"
3. as a notice from a local traffic police's blog.

We argue that Twitter application does not provide enough factors to evaluate the reliability^{p} of the information creation process, that is the reliability^{p} of the person X to produce the piece of information.[14] Therefore, ISA$_{bdi}$-A does not declare even a traffic notice based on this information. The same applies to the second case, as well.

Conclusion: In this case our intuition says that ISA$_{bdi}$-A is dependable.

The third case is more interesting, because the reliability^{p} of the processes of the traffic police (the traffic police is a well–organized public organization) can be evaluated, and the police blog application may provide metadata about the reliability^{p}. If there is no such metadata, then ISA$_{bdi}$-A could carry out a reliability^{p} assessment for the piece of information using, for example, the following factors: social media application, source, author position, and reputation. Reputation comprises the following factors: past history related to the content creation, responses to the previous content, and generic interaction with others. As such this

14 For example, there is possibility of a fake profile or the Twitter account might haven cracked.

case resembles the case 6 above. Based on the result of the warrant, ISA$_{bdi}$-A declares the traffic alert both to ADC and HDC.

Conclusion: In this case we consider ISA$_{bdi}$-A to be dependable. When ISA$_{bdi}$-A operates on beliefs, which is perceived from social media, ISA$_{bdi}$-A shall itself evaluate or request a warrant service to carry out the evaluation of each piece of information.

The key issues are as follows:

1. Theory: reliabilism and testimony.
2. The epistemic quality of information in various social media.
3. The reliability^{p} of unknown or uncertified human being to distribute information in various social media.
4. The reliability^{p} of the processes of the social media to distribute information (cracking problems).
5. In reality the piece of information may represent knowledge but produced via a not–reliable^{p} enough process, which causes undependable system.

3.8 Common Information Service

In this case (illustrated in Figure 13) we assume that ISA$_{bdi}$-A perceives the piece of information stated by the proposition *"Road 101 is slippery."* from a common information service.

There are several different kinds of common information services, such as healthcare information services (e.g. medical records by Kela[15] and Terveyskirjasto by Duodecim[16]), administrative information services (e.g. Suomi.fi by Finnish government[17]), traffic security information services (e.g. Trafi[18]), domestic security information services (e.g. poliisi[19]), and airport information services (e.g. Finavia[20]).

15 www.kanta.fi
16 www.terveyskirjasto.fi
17 www.suomi.f
18 www.trafi.fi
19 www.poliisi.fi
20 https://www.finavia.fi

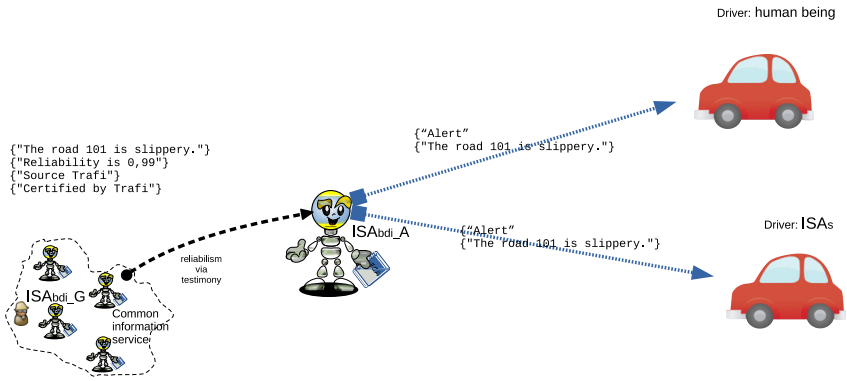

Figure 13: Common Information Service.

There are two possibilities: first, the reliability^{p} of the information creation processes of the common information service has not been evaluated, and second, the reliability^{p} has been evaluated (for example, by a governmental authority). Let us assume that in the first case ISA$_{bdi}$-A perceives from a common information service *"Road 101 is slippery."* with associated metadata *"Time 02.03.2020 14:00" "Source Association of Taxi Drivers"*. There is no data about the reliability^{p} of the Association of Taxi Drivers–produced information. Thus, the epistemic quality is unknown, and ISA$_{bdi}$-A does not declare the traffic notice, warning or alert.

Conclusion: In this case our intuition says that ISA$_{bdi}$-A is dependable.

In the second case, ISA$_{bdi}$-A perceives from the common information service "Road 101 is slippery." with associated metadata *"Reliability^{p} is 0.99" "Source Trafi" "Warranted by Trafi"* (Figure 13). Our intuition says that the process of Trafi (based on actions during long history) is reliable^{p} enough (without a warrant done by

another government party) in the domain of traffic services to produce belief, which creation processes fulfil the reliability^{p} requirements for knowledge. Therefore, we consider that the belief *"Road 101 is slippery."* is knowledge, and ISA$_{bdi}$-A declares the traffic alert.

Conclusion: In this case we consider ISA$_{bdi}$-A to be dependable.

The key issues are as follows:

1. Theory: reliabilism and reliabilism via testimony.
2. The reliability^{p} of various common services to produce information.
3. The evaluation or warrant of common services.
4. Can a common service be trusted a priori to provide the reliability^{p} data?
5. A belief without any data about the reliability^{p} of the belief creation process shall be neglected.
6. Does an individual inherit the reliability^{p} of processes of an organization?

3.9 Another ISA$_{bdi}$-X

In this case (illustrated in Figure 14) we assume that ISA$_{bdi}$-A perceives the piece of information stated by the proposition *"Road 101 is slippery."* from another ISA$_{bdi}$-X.

There are four different cases as listed in Table 3: First, a certified ISA$_{bdi}$-X operates on beliefs with embedded metadata providing the reliability^{p} of the belief forming process (and the reliability^{p} of the justification forming process). Second, the certified ISA$_{bdi}$-X operates on beliefs without any data about the reliability^{p} of the belief forming process. Third, the non–certified ISA$_{bdi}$-X operates on beliefs with embedded metadata providing the reliability^{p} of the belief forming process (and the reliability^{p} of the justification forming process). Fourth, the non–certified ISA$_{bdi}$-X operates on beliefs without any data about the reliability^{p} of the belief forming process.

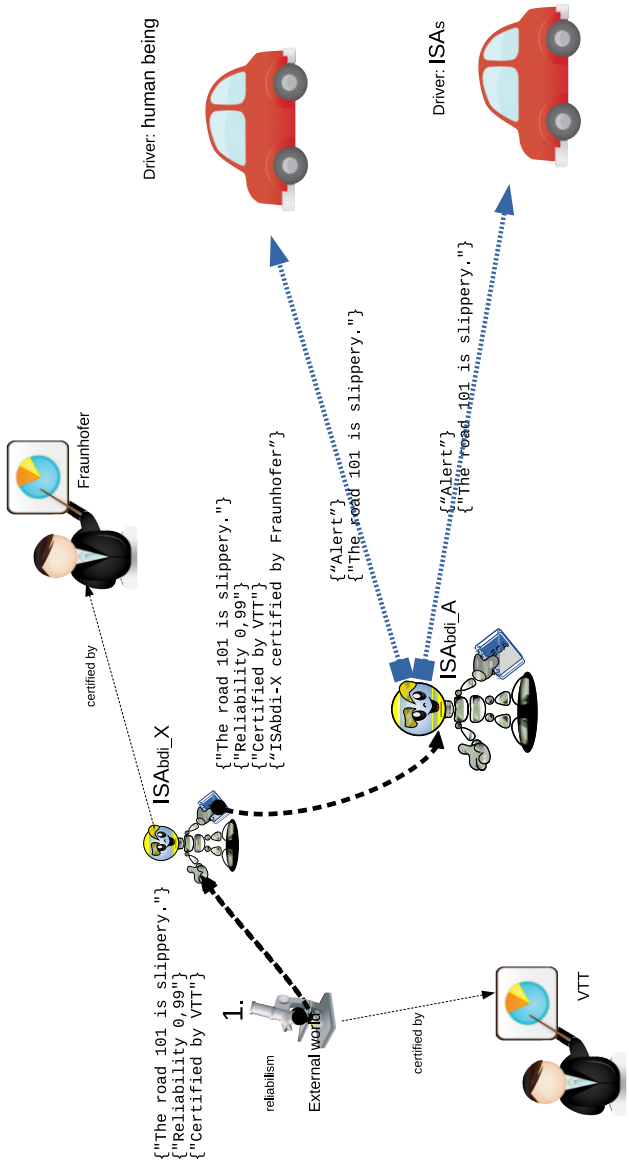

Figure 14: Another ISA.

1	2
ISA$_{bdi}$-X certified	**ISA$_{bdi}$-X certified**
Reliability of belief known	Reliability of belief not known
3	4
ISA$_{bdi}$-X not certified	**ISA$_{bdi}$-X not certified**
Reliability of belief known	Reliability of belief not known

Table 3: Cases of ISA$_{bdi}$ as source of information.

The first case is an example of testimonially transferred belief, therefore the key question is the one whether the justification in the metadata is well–grounded for the receiver of the belief to be justified to believe or to know the belief. Is there a requirement for additional justification? The second, third, and fourth cases demand to evaluate the reliability^{p} of the belief forming process, for example, using a warrant service. In other words, the evaluation of the reliability^{p} of the ISA$_{bdi}$ or the reliability of the warrant service to form or certify such a belief. There are several factors, which can be used in the evaluation:

1. The manufacturer of the ISA$_{bdi}$
2. The accuracy of the ISA$_{bdi}$ in past actions
3. The appropriateness of the ISA$_{bdi}$ to issue the belief
4. The consistency of the ISA$_{bdi}$ in past actions
5. The reputation of the ISA$_{bdi}$ (history)
6. The timeliness of the belief.

We argue that ISA$_{bdi}$ as a source of belief is comparable to a human being. Therefore, the case of the human being (Section 2.6) is applicable in this case.

The key issues are as follows:

1. Theory: reliabilism and testimony.
2. The reliability^{p} of ISA$_{bdi}$ to produce belief.
3. The evaluation or warrant of ISA$_{bdi}$.

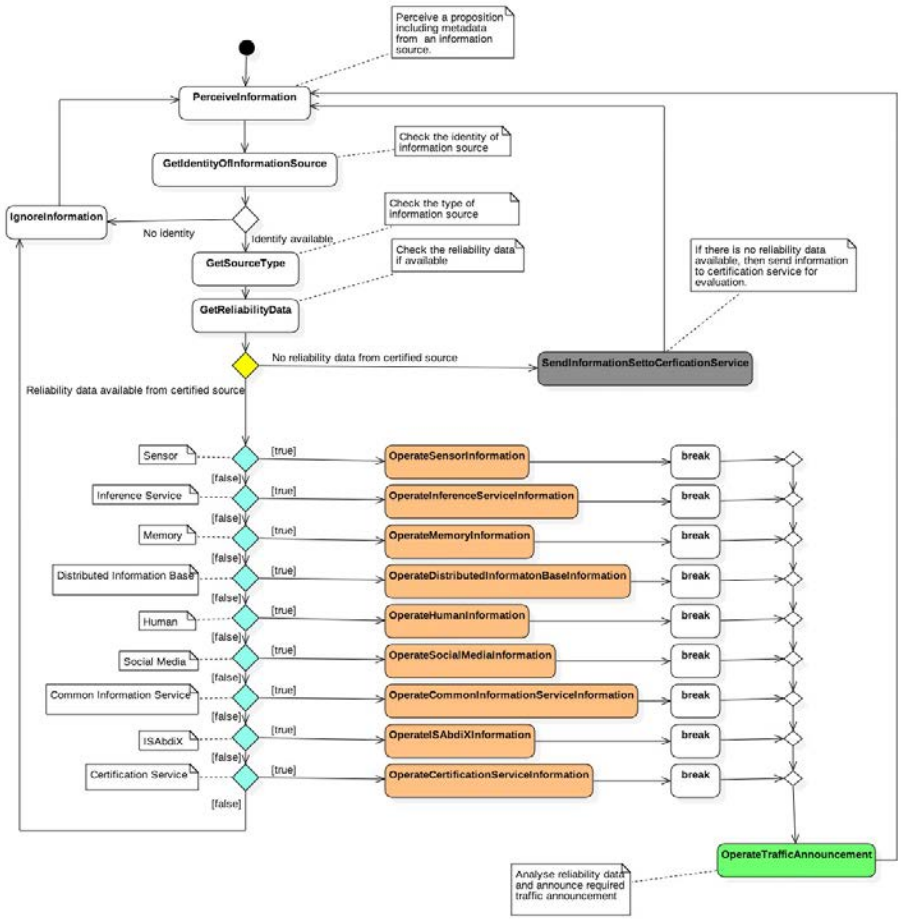

Figure 15: Activities of Traffic Announcement.

4. The warrant service is the solution in the case of $\mathrm{ISA}_{\mathtt{bdi}}$ (either beforehand or in real time).

5. The warrant of an individual $\mathrm{ISA}_{\mathtt{bdi}}$ is a difficult issue. Can it be based on the same principles as the warrant of a human being.

6. Does an individual $\mathrm{ISA}_{\mathtt{bdi}}$ inherit the reliability^{p} of the processes of a whole multi-agent system?

We can summarize the above discussion in the form of a UML activity diagram as illustrated in Figure 15.

4 ONTOLOGY OF TRAFFIC INFORMATION SYSTEM

A simplified ontology of this scenario is illustrated in Figures 16 and 17.

A more detailed ontology can be found in the appendix and in more readable form on http://www.heimolaamanen.fi .

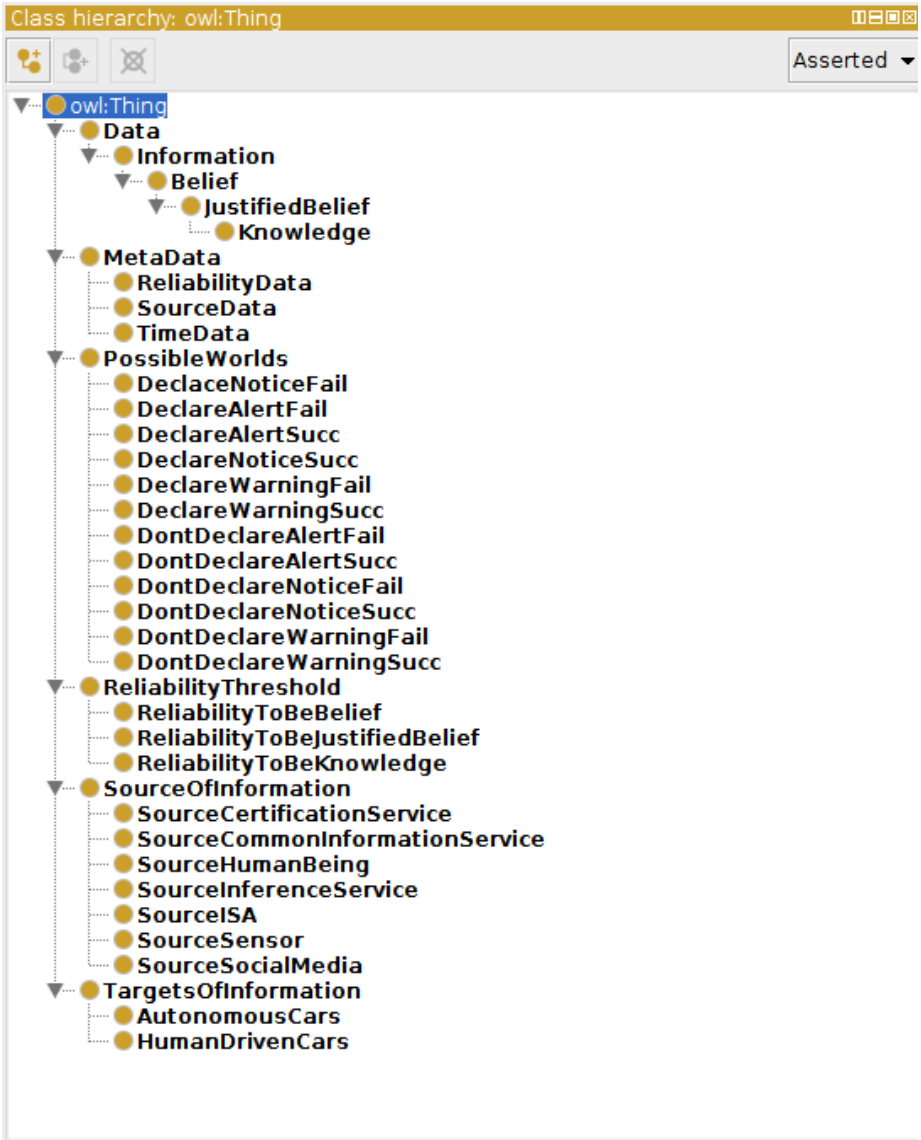

Figure 16: Class hierarchy.

Figure 17: Object property hierarchy.

5 SUMMARY

We claimed that in the future intelligent software agents will produce highly sophisticated services in co–operation (having equal operational status) with human beings. This requires that ISA_{bdi}s should have similar concepts about the epistemological quality of information with human beings. In this paper we discussed several issues related to the epistemological quality of information in the contexts of various information sources. We showed that the epistemological quality of information affect the dependability of intelligent software agents (actually in a similar way as it affects the dependability of human beings).

We summarize our discussion in Table 4.

♯	Source	Theory	General Features
1	Sensor	Reliabilism	Reliability available—either manufacturer provide it or it can be tested.
2	Inference service	Realiabilism	Reliability available—can be evaluated from the reliability of different sources and inferring methods.
3	Memory	Reliabilism	Reliability available—can be evaluated from the reliability of different sources. Up–to–datedness.
4	Distributed information base	Reliabilism	Reliability available—can be evaluated from the reliability of different sources. Up–to–dateness.
5	Warrant service	Testimony Reliabilism	Reliability available by third party—in addition requires reliability of justification. Autonomous, automatic warrant services?
6	Human being	Testimony	Reliability is not available — requires use of warrant service or certificated human being.
7	Social media	Testimony	Reliability is not available—requires use of warrant service or new social media application, which can manage demanded metadata and their creation.
8	Common information service	Testimony Reliabilism	Reliability either available or not available depending on the status of the service. Certified service provider.
9	Another $\mathrm{ISA}_{\mathtt{bdi}}$	Testimony Reliabilism	Reliability either available or not available depending on the $\mathrm{ISA}_{\mathtt{bdi}}$-X. May require use of warrant service or certificated $\mathrm{ISA}_{\mathtt{bdi}}$.

Table 4: Summary of Sources of Belief.

Appendices

A ONTOLOGY

Classes

AutonomousCars

AutonomousCars $\sqsubseteq$ TargetsOfInformation
AutonomousCars $\sqsubseteq \neg$ HumanDrivenCars

Belief

Belief $\sqsubseteq$ Information

Data

Data $\sqsubseteq \neg$ PossibleWorlds

DeclaceNoticeFail

DeclaceNoticeFail $\sqsubseteq$ PossibleWorlds

DeclareAlertFail

DeclareAlertFail $\sqsubseteq$ PossibleWorlds

DeclareAlertSucc

DeclareAlertSucc $\sqsubseteq$ PossibleWorlds

DeclareNoticeSucc

DeclareNoticeSucc $\sqsubseteq$ PossibleWorlds

DeclareWarningFail

DeclareWarningFail $\sqsubseteq$ PossibleWorlds

DeclareWarningSucc

DeclareWarningSucc $\sqsubseteq$ PossibleWorlds

DontDeclareAlertFail

DontDeclareAlertFail $\sqsubseteq$ PossibleWorlds

DontDeclareAlertSucc

DontDeclareAlertSucc $\sqsubseteq$ PossibleWorlds

DontDeclareNoticeFail

DontDeclareNoticeFail $\sqsubseteq$ PossibleWorlds

DontDeclareNoticeSucc

DontDeclareNoticeSucc $\sqsubseteq$ PossibleWorlds

DontDeclareWarningFail

DontDeclareWarningFail $\sqsubseteq$ PossibleWorlds

DontDeclareWarningSucc

DontDeclareWarningSucc $\sqsubseteq$ PossibleWorlds

HumanDrivenCars

HumanDrivenCars $\sqsubseteq$ TargetsOfInformation
HumanDrivenCars $\sqsubseteq \neg$ AutonomousCars

Information

Information $\sqsubseteq$ Data

JustifiedBelief

JustifiedBelief $\sqsubseteq$ Belief

Knowledge

Knowledge $\sqsubseteq$ JustifiedBelief

MetaData

PossibleWorlds

PossibleWorlds $\sqsubseteq \neg$ SourceOfInformation
PossibleWorlds $\sqsubseteq \neg$ Data
PossibleWorlds $\sqsubseteq \neg$ ReliabilityThreshold

ReliabilityData

ReliabilityData $\sqsubseteq$ MetaData

ReliabilityThreshold

ReliabilityThreshold $\sqsubseteq \neg$ PossibleWorlds

ReliabilityToBeBelief

ReliabilityToBeBelief $\sqsubseteq$ ReliabilityThreshold

ReliabilityToBeJustifiedBelief

ReliabilityToBeJustifiedBelief $\sqsubseteq$ ReliabilityThreshold

ReliabilityToBeKnowledge

ReliabilityToBeKnowledge $\sqsubseteq$ ReliabilityThreshold

SourceCertificationService

SourceCertificationService $\sqsubseteq$ SourceOfInformation

SourceCommonInformationService

SourceCommonInformationService $\sqsubseteq$ SourceOfInformation

SourceData

SourceData $\sqsubseteq$ MetaData

SourceHumanBeing

SourceHumanBeing $\sqsubseteq$ SourceOfInformation

SourceISA

SourceISA $\sqsubseteq$ SourceOfInformation

SourceInferenceService

SourceInferenceService $\sqsubseteq$ SourceOfInformation

SourceOfInformation

SourceOfInformation $\sqsubseteq \neg$ PossibleWorlds

SourceSensor

SourceSensor $\sqsubseteq$ SourceOfInformation

SourceSocialMedia

SourceSocialMedia $\sqsubseteq$ SourceOfInformation

TargetsOfInformation

TimeData

TimeData $\sqsubseteq$ MetaData

OBJECT PROPERTIES

hasAffectOn

$\exists$ hasAffectOn Thing $\sqsubseteq \exists$ hasAffectOn PossibleWorlds
$\top \sqsubseteq \forall$ hasAffectOn ($\exists$ hasAffectOn TargetsOfInformation)

hasMetaData

$\exists$ hasMetaData Thing $\sqsubseteq \exists$ hasMetaData SourceSocialMedia
$\exists$ hasMetaData Thing $\sqsubseteq \exists$ hasMetaData SourceCertificationService
$\exists$ hasMetaData Thing $\sqsubseteq \exists$ hasMetaData SourceCommonInformationService
$\exists$ hasMetaData Thing $\sqsubseteq \exists$ hasMetaData SourceHumanBeing
$\exists$ hasMetaData Thing $\sqsubseteq \exists$ hasMetaData SourceISA
$\exists$ hasMetaData Thing $\sqsubseteq \exists$ hasMetaData SourceInferenceService

∃ hasMetaData Thing ⊑ ∃ hasMetaData SourceSensor
⊤ ⊑ ∀ hasMetaData (∃ hasMetaData MetaData)

hasReliabilityThreshold

hasReliabilityToBeBelief

⊑ hasReliabilityThreshold
∃ hasReliabilityToBeBelief Thing ⊑ ∃ hasReliabilityToBeBelief Belief
⊤ ⊑ ∀ hasReliabilityToBeBelief (∃ hasReliabilityToBeBelief ReliabilityToBeBelief)

hasReliabilityToBeJustifiedBelief

⊑ hasReliabilityThreshold
∃ hasReliabilityToBeJustifiedBelief Thing ⊑ ∃ hasReliabilityToBeBelief JustifiedBelief
⊤ ⊑ ∀ hasReliabilityToBeJustifiedBelief (∃ hasReliabilityToBeJustifiedBelief ReliabilityToBeJustifiedBelief)

hasReliabilityToBeKnowledge

⊑ hasReliabilityThreshold
∃ hasReliabilityToBeKnowledge Thing ⊑ ∃ hasReliabilityToBeKnowledge Knowledge
⊤ ⊑ ∀ hasReliabilityToBeKnowledge (∃ hasReliabilityToBeKnowledge ReliabilityToBeKnowledge)

hasSource

hasValueFrom

∃ hasValueFrom Thing ⊑ ∃ hasValueFrom ReliabilityToBeJustifiedBelief
∃ hasValueFrom Thing ⊑ ∃ hasValueFrom ReliabilityToBeKnowledge
∃ hasValueFrom Thing ⊑ ∃ hasValueFrom ReliabilityToBeBelief
⊤ ⊑ ∀ hasValueFrom (∃ hasValueFrom PossibleWorlds)

DATA PROPERTIES

ReliabilityLimit

Value

INDIVIDUALS

ADCReliabilityLimitDAF

ADCReliabilityLimitDAF : DeclareAlertFail
  ReliabilityLimit (ADCReliabilityLimitDAF "0.95"
      http://www.w3.org/2001/XMLSchema)

ADCReliabilityLimitDAS

ADCReliabilityLimitDAS : DeclareAlertSucc
  ReliabilityLimit (ADCReliabilityLimitDAS "0.95"
      http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDDAF

ADCReliabilityLimitDDAF : DontDeclareAlertFail
  ReliabilityLimit (ADCReliabilityLimitDDAF "0.01"
      http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDDAS

ADCReliabilityLimitDDAS : DontDeclareAlertSucc
  ReliabilityLimit (ADCReliabilityLimitDDAS "0.99"
      http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDDNF

ADCReliabilityLimitDDNF : DontDeclareNoticeFail
  ReliabilityLimit (ADCReliabilityLimitDDNF "0.30"
      http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDDNS

ADCReliabilityLimitDDNS : DontDeclareNoticeSucc
  ReliabilityLimit (ADCReliabilityLimitDDNS "0.70"
      http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDDWF

ADCReliabilityLimitDDWF : DontDeclareWarningFail
  ReliabilityLimit (ADCReliabilityLimitDDWF "0.15"
    http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDDWS

ADCReliabilityLimitDDWS : DontDeclareWarningSucc
  ReliabilityLimit (ADCReliabilityLimitDDWS "0.85"
    http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDNF

ADCReliabilityLimitDNF : DeclaceNoticeFail
  ReliabilityLimit (ADCReliabilityLimitDNF "0.35"
http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDNS

ADCReliabilityLimitDNS : DeclareNoticeSucc
  ReliabilityLimit (ADCReliabilityLimitDNS "0.65"
    http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDWF

ADCReliabilityLimitDWF : DeclareWarningFail
  ReliabilityLimit (ADCReliabilityLimitDWF "0.20"
    http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityLimitDWS

ADCReliabilityLimitDWS : DeclareWarningSucc
  ReliabilityLimit (ADCReliabilityLimitDWS "0.80"
    http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityToBeBelief

ADCReliabilityToBeBelief : ReliabilityToBeBelief

ADCReliabilityToBeBelief-1

ReliabilityLimit (ADCReliabilityToBeBelief-1 "0.70"
    http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityToBeJustifiedBelief

ADCReliabilityToBeJustifiedBelief : ReliabilityToBeJustifiedBelief
  ReliabilityLimit (ADCReliabilityToBeJustifiedBelief "0.85"
      http://www.w3.org/2001/XMLSchema♯decimal)

ADCReliabilityToBeKnowledge

ADCReliabilityToBeKnowledge : ReliabilityToBeKnowledge
  ReliabilityLimit (ADCReliabilityToBeKnowledge "0.99"
      http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDAF

HDCReliabilityLimitDAF : DeclareAlertFail
  ReliabilityLimit (HDCReliabilityLimitDAF "0.10"
      http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDAS

HDCReliabilityLimitDAS : DeclareAlertSucc
  ReliabilityLimit (HDCReliabilityLimitDAS "0.90"
      http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDDAF

HDCReliabilityLimitDDAF : DontDeclareAlertFail
  ReliabilityLimit (HDCReliabilityLimitDDAF "0.05"
      http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDDAS

HDCReliabilityLimitDDAS : DontDeclareAlertSucc
  ReliabilityLimit (HDCReliabilityLimitDDAS "0.95"
      http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDDNF

HDCReliabilityLimitDDNF : DontDeclareNoticeFail
  ReliabilityLimit (HDCReliabilityLimitDDNF "0.35"
      http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDDNS

HDCReliabilityLimitDDNS : DontDeclareNoticeSucc
ReliabilityLimit (HDCReliabilityLimitDDNS "0.65"
http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDDWF

HDCReliabilityLimitDDWF : DontDeclareWarningFail
ReliabilityLimit (HDCReliabilityLimitDDWF "0.23"
http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDDWS

HDCReliabilityLimitDDWS : DontDeclareWarningSucc
ReliabilityLimit (HDCReliabilityLimitDDWS "0.77"
http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDNF

HDCReliabilityLimitDNF : DeclaceNoticeFail
ReliabilityLimit (HDCReliabilityLimitDNF "0.40"
http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDNS

HDCReliabilityLimitDNS : DeclareNoticeSucc
ReliabilityLimit (HDCReliabilityLimitDNS "0.60"
http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDWF

HDCReliabilityLimitDWF : DeclareWarningFail
ReliabilityLimit (HDCReliabilityLimitDWF "0.26"
http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityLimitDWS

HDCReliabilityLimitDWS : DeclareWarningSucc
ReliabilityLimit (HDCReliabilityLimitDWS "0.74"
http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityToBeBelief

HDCReliabilityToBeBelief : ReliabilityToBeBelief
  ReliabilityLimit (HDCReliabilityToBeBelief "0.65"
      http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityToBeJustifiedBelief

HDCReliabilityToBeJustifiedBelief : ReliabilityToBeJustifiedBelief
  ReliabilityLimit (HDCReliabilityToBeJustifiedBelief "0.77"
      http://www.w3.org/2001/XMLSchema♯decimal)

HDCReliabilityToBeKnowledge

HDCReliabilityToBeKnowledge : ReliabilityToBeKnowledge
  ReliabilityLimit (HDCReliabilityToBeKnowledge "0.95"
      http://www.w3.org/2001/XMLSchema♯decimal)

DATATYPES

PlainLiteral

decimal

REFERENCES

[1] Alvin I. Goldman. Internalism exposed. In Ernest Sosa, Jaegwon Kim, Jeremy Fantl, and Matthew McGrath, editors, *Epistemology: An Anthology*, chapter Internalism Exposed, page 379–393. Blackwell Publishing, second edition, 2008. ISBN 978-1-4051-6967-7.

[2] Alvin I. Goldman. What is justified belief? In Ernest Sosa, Jaegwon, Jeremy Fantl, and Matthew McGrath, editors, *Epistemology: An Anthology*, chapter What is Justified Belief?, page 333–347. Blackwell Publishing, second edition, 2008. ISBN 978-1-4051-6967-7.

[3] Contributors of Wikipedia. Wikipedia. In *Fact checker*. Wikipedia, The Free Encyclopedia, 2016.

[4] Giovanni Ciampaglia, Prashant Shiralkar, Luis Rocha, Johan Bollen, Filippo Menczer, and Allessandro Flammini. Computational fact checking from knowledge networks. *PLOS*, 10(10), June 2015. Available on http://journals.plos.org/article?id=10.1371-/journal.pone.0128193.

[5] Jason Nurse, Syed Rahman Sadie aand Creese, Michael Goldsmith, and Koen Lamberts. Information quality and trustworthiness: A topical state-of-the-art review. In *2011 International Conference on Computer Applications and Network Security (ICCANS 2011)*. IEEE, 2011.

[6] Jason Nurse, Ioannis Agrafiotis, Michael Goldsmith, Sadie Creese, Koen Lamberts, Darren Price, and Glyn Jones. Information trustworthiness as a solution to the misinformation problems in social media. In *1st Internation Conference on Cyber Security for Sustainable Society 2015*, page 28–35, 2015. ISSN 2052-8604.

[7] Sai T. Moturu and Huan Liu. Quantifying the trustworthiness of social media content. *Distributed and Parallel Databases*, 29(3):239–260, 2011.

[8] Jason Nurse, Ioannis Agrafiotis, Michael Goldsmith, Sadie Creese, and Koen Lamberts. Two sides of the coin: measur-

ing and communicating the trustworthiness of online information. *Journal of Trust Management*, 1:5, 5 2014. Available on journaloftrustmanagement.springeropen.com/articles-/10.1186/2196-064.

[9] O. Arazy and R. Kopak. On the measurability of information quality. *American Society for Information Science and Technology (JASIST)*, 62(1):89–99, 2011.

[10] Mona Alkhattabi, Daniel Neagu, and Andrea Cullen. Information quality framework for e-learning systems. *Knowledge Management & E-Learning: An International Journal*, 2(4):340–362, 2010. Available on http://scim.brad.ac.uk/staff/pdf/ajcullen/21-227-2-PB.pdf.

[11] Diane Strong, Yang W. Lee, and Richard Y. Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, May 1997.

[12] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *WWW'11 Proceedings of the 20th Internation Conference on World Wide Web*, page 675–684. ACM, New York, NY, USA, 2011.

[13] Syed Rahman, Sadie Creese, and Michael Goldsmith. Accepting information with a pinch of salt: Hanling untrusted information sources. *Security and Trust Management*, page 223–238, 2012.

APPENDIX 4

—

IS IT TIME TO GET OUT OF THE CHINESE ROOM?

CONTENTS

1 The Chinese Room Argument 1
2 Replies to the Chinese Room Argument 5
2.1 The Systems Reply 5
2.2 The Virtual Mind Reply 6
2.3 The Robot Reply 6
2.4 The Brain Simulator Reply 7
2.5 The Other Minds Reply 8
2.6 The Intuition Reply 8
3 Issues on Syntax and Semantics 9
4 Issues about Intentionality and Consciousness 12
5 Some Other Thoughts 13
6 Summary 14

1 THE CHINESE ROOM ARGUMENT

John R. Searle seems to strongly believe that homo sapiens—having the ability to make and use complex tools—is one of the kind, and there will never be a digital computer (an artificial entity), which capabilities (intentionality, understanding, qualia) are equal or exceeds the ones of human beings. He has tried to prove his argument and refute the Turing Test[1] with a thought experiment called Chinese Room (hereinafter CRA). The CRA argument is the following one [2][1] *Suppose that a man is locked in a room and given a large batch of Chinese writing. Suppose furthermore that the man knows no Chinese, either written or spoken, and that the man is not even confident that he could not recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. Now suppose further that after this first batch of Chinese writing the man is given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and the man understands these rules as well as any other native speaker of English. They enable the man to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that the man can identify the symbols entirely by their shapes. Now suppose also that the man is given a third batch of Chinese symbols together with some instruction, again in English, that enable the man to correlate elements of this third batch with the first two batches, and these rules instruct the man how to give back certain Chines symbols with certain sorts of shapes given the man in the third batch. Unknown to the man, the people who are giving the man all of these symbols call the first batch "a script", they call the second batch "a story", and they call the third batch "questions". Furthermore, they call the symbols the man gives them back in response to third batch "answers to the questions", and the set of rules in English that they gave the man, they call "the program". Now just complicate the story a little, image that these also give stories in English, which the man understands, and they then ask the man questions in English about these stories, and the man gives them back answers in English. Suppose also that after a while the man gets so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from the point of view of somebody outside*

1 The word "I" has been replaced with the word "man" in the text below.

the room in which the man locked—the man's answers to the questions are absolutely indistinguishable from the native Chinese speakers. Nobody just looking at the man's answers can tell that the man doesn't speak a word of Chinese. Let us also supposes that the man's answers to the English questions are, as they no doubt would be, indistinguishable from those of a native English speaker. From the external point of view—from the point of view of someone reading the man's "answers"—the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, the man produces the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, the man simply behaves like a computer; the man performs computational operations on formally specified elements. For the purpose of the Chinese, the man is simply an instantiation of the computer program.

Searle himself summarized his claims in the abstract of his article *Minds, brains, and programs* as follows [2]:

1. *"Intentionality in human beings (and animals) is a product of causal features of the brain. I assume this is an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are sufficient for intentionality. "*

2. *"Instantiating a computer program is never by itself a sufficient condition of intentionality. The main argument is directed at establishing this claim. The form of the argument is to show how a human agent could instantiate the program and still not have the relevant intentionality. "*

These two propositions have the following consequences:

- *"The explanation of how the brain produces intentionality cannot be that it does it by instantiating a computer program. This is a strict logical consequence of 1 and 2. "*

- *"Any mechanism capable of producing intentionality must have causal powers equal to those of the brain. This is meant to be a trivial consequence of 1."*

3. *"Any attempt literally to create intentionality artificially (strong AI) could not succeed just by designing programs but would have to duplicate the causal powers of the human brain."*

The programs implemented by computers are just syntactical, and they do not responds to the meaning of symbols as minds do. The main target of the CRA was to argue that there is no artificial entities based on digital computers having any genuine psychological properties solely in virtue of its running a program [3]. Searle's aim was also to refute the functionalist approach to understand minds [4]. This is because of that the room not only behaves as if it understood Chinese, it functions as if it does [3].

Searle's argument, that because syntax is not sufficient for semantics, programs cannot produce mind, is the following one [3]:

1. Programs are purely formal (syntactic).
2. Human minds have mental contents (semantics).
3. Syntax by itself is neither constitutive of, nor sufficient for, semantic content.
4. Therefore, programs by themselves are not constitutive of nor sufficient for minds.

Later in 2002 Searle expressed that his fundamental claim as follows [5]: *"The purely formal or abstract or syntactical processes of the implemented computer program could not by themselves be sufficient to guarantee the presence of mental content or semantic content of the sort that is essential to human cognition. Of course a system might have semantic content for some other reason, but it does not apply to Strong Artificial Intelligence, any more. The basic structure of the Chinese Room Argument is rather obvious: the distinction between syntax and semantics and the distinction between simulation and duplication."*

In this statement Searle admits that a system might have semantic content for another reason than Strong Artificial Intelligence (hereinafter SAI). Then what is the SAI? There is no unambiguous definition. It seems to be a term, which is used to describe a particular mindset of Artificial Intelligence (hereinafter AI) such as the goal of the SAI is to develop AI to the level where the machine's intellectual capabilities is functionality equal to ones of human beings. Stuart Russell and Peter Norvig have defined the SAI as follows [6]: *The assertion that machines do actually thinking as opposed to simulate thinking*. Searle himself has defined the SAI to be the following one: *The appropriately programmed computer really is a mind, in the sense that computers given the right programs can literally said to understand and have other cognitive states*

[2]. Thus, there is a difference between simulating understanding (thinking) (Weak Artificial Intelligence) and really understanding (thinking).[2] Searle's definition beg questions: What does he mean with *"literally said to understand"*? What is the difference between real understanding and simulated understanding?

Searle [5] claims that the SAI is a weird mixture of behaviourism and dualism. It is behaviourist in its acceptance of the Turing Test and dualist by rejecting the idea that consciousness and intentionality are ordinary biological phenomena like digestion. If you accept the combination of behaviourism and dualism, then it is natural to think that the mind is a substantive physical process, and it is something formal and abstract.

It is an open question what sorts of systems are necessary and sufficient to produce consciousness and intentionality, and it is likely to remain open until we figure out how brains do it. In order to create consciousness you have to create mechanism which can duplicate and not merely simulate the capacity of the brain to create consciousness [5]. Thus, Searle argues that in this case there is a substantial difference between simulation and duplication.

The CRA has raised a lot of writings and published articles discussing of the possibility of digital computers to understand language or think. In the late 70's and earlier 80's, at the time when John R. Searle wrote his article, AI was experiencing its first real hype; there was no limits what AI could do in the future. But it turned out that the development and implementation of algorithms required to fulfil the promises of AI were at that time too difficult, because of the inadequate state of computer science and low performance of computers. AI [3] was seen to be an attempt to design and build computer systems, which display a range of genuine psychological attributes such as problem-solving, thinking, understanding, and reasoning. In addition, an ultimate goal AI was to produce consciousness, feeling, and emotion. Thus, mental processes are seen to be computational processes over formally defined elements. We need to regard the CRA in this background.

Is Searle correct when claiming that a digital computer is just a device that manipulates symbols without a possibility to have semantics, understanding, or consciousness? AI and the perfor-

2 As the CRA is about understanding, I mainly concentrate on understanding and not on other properties of intentionality.

mance of computers have progressed significantly during the last 30 years, and this sets the CRA into a more challenging framework. At the moment we are experiencing the hype of AI, once again. For example, a computer applications based on AI has won best human players on television quiz show Jeopardy[3] indicating understanding of natural languages and stories, and robots are shown to provide elderly people with healthcare and social care, or develop their own language and teach it to other robots indicating to have mental contents [7]. I will evaluate the CRA in this new framework in this essay, but at first I present the major replies to the Chinese Room Argument.

2 REPLIES TO THE CHINESE ROOM ARGUMENT

There are several different approaches to reply to the CRA. The main lines are the following ones [4]:

- The man in the room does not understand Chinese, but executing the program (as a whole) may establish something that understands Chinese.
- A different kind of a computer system (e.g. robot) could understand Chinese.
- The scenario is not feasible and our intuitions in such cases are unreliable. For example, what does Searle mean by *"understand"* and *"intentionality"*?

2.1 The Systems Reply

Searle has said that the Systems Reply is the most common one. According to the Systems Reply the man in the room does not understand English, but the man is just a part in a larger system, and the larger system as a whole understands Chinese. Several philosophers (Ned Block, Daniel Dennett, Jerry Fodor, John Haugeland, and Ray Kurzweil) have supported the System Reply.

3 www.jeopardy.com

Searle has responded to the Systems Reply by stating that even though the man can internalize the whole system and leave the room and wander outdoors conversing Chinese, the man still would have no way to attach *"any meaning to the formal symbols"*. The man would be the entire system, yet he still would not understand Chinese. The System has no more means to attach semantics to the Chinese symbols than the person in the room has, because a system can have no psychological properties not possessed by its subsystems [3].

2.2 The Virtual Mind Reply

The key idea of the Virtual Mind Reply is that even though the man in the room does not himself understand Chinese, the important thing is that **understanding is created**. An active system may create new, virtual, entities that are distinct from the system as a whole or the subsystems (e.g. central processing unit and memory). A virtual entity may create an agent that understand Chinese (as an example, Apple's Siri understands English reasonably well) [4]. The CRA is wrong in claiming that the SAI is about *"the computer understand Chinese"* or *"the System understands Chinese"*, but the appropriate issue for the SAI is whether *"the running computer creates understanding of Chinese"*. We should distinguish between minds and their realizing systems. The CRA cannot controvert the following the SAI claim: it is possible to create understanding using a programmed digital computer [4]. This kind of approach is supported by following philosophers: Marvin Minsky, Aaron Sloman, Monica Croucher, David Chalmers, and Ned Block. Searle has not actually responded properly to the Virtual Mind Reply.

2.3 The Robot Reply

The Robot Reply [4] agrees with Searle that a computer in a computer room cannot understand a language, or know what words mean. Understanding and knowing require proper connections to the external world. Therefore, the Robot Reply suggests that a digital computer is put into a robot body provided with required perception equipment, and effectors equipment to move around

and to manipulate its environment. This kind of a robot could learn by seeing and doing. It can attach meanings to symbols, and thus actually understand a natural language. This approach supports externalist semantics by stating that suitable causal connections with the world can give a semantic meaning to internal symbols. Therefore, a robot can have propositional attitudes. Tim Crane [4] argues that *"The proper response to Searle's argument is: Sure, Searle–in–the–room, or the room alone, cannot understand Chinese. But if you let the outside world have some impact on the room, meaning or 'semantics' might begin to get a foothold. But of course, this concedes that thinking cannot be simply symbol manipulation."* . Daniel Dennett, Jerry Fodor, and Georges Rey have supported the Robot Reply.

Searle considers that the Robot Reply to the CRA is not any better than the Systems Reply, because sensors just provide additional (only syntactic) input to the computer. Searle argues that this syntactic input will do nothing to allow the man to associate meanings with Chinese characters [4]. Searle claims that there is a wrong level of causation. The robot Reply leaves out the normative dimension of intentional concepts, as intentionality is irreducible [3]. Stevan Harnad argues that feelings—such as the feeling of understanding—are missing [8].

2.4 The Brain Simulator Reply

The Brain Simulator Reply supposes that a computer simulates the actual sequences of nerve operations that occur in the brain—connectionism—of a native Chinese language speaker when that person understands Chinese. Since the computer operates the very same way as the brain, it will understand Chinese. Paul and Patricia Churchland are proponents of this approach.

According to Searle, simulation does not make any difference, because a simulation of brain activity is not the real thing [4]. Searle agrees that it would be reasonable to attribute understanding to an android system (totally human–like artificial entity), but only as long as you don't know how it works. As soon as you know the truth—it is a computer manipulating symbols on the basis of syntax, not meaning–you would cease to attribute inten-

tionality to it [2]. The computational power of neural networks is no stronger than that Turing machines.

2.5 The Other Minds Reply

The key point of the Other Mind Reply is the following one: *"How do you know that other people understand Chinese or anything else? Only by their behaviour. Now the computer can pass the behavioural tests as well as they can (in principle), so if you are going to attribute cognition to other people you must in principle also attribute it to computers."* [4]. Now, Searle's reply to this approach contains a very interesting point [2]: *The problem in this discussion is not about how I know that other people have cognitive states, but rather what it is that* ***I am attributing to them when I attribute cognitive states to them****. The thrust of the argument is that it couldn't be just computational processes and their output because the computational processes and their output can exist without the cognitive state.* This statement raises the question: Is Searle attributing same properties as other people involved in this discussion do?

2.6 The Intuition Reply

According to the Intuition Reply the CRA seems to be based on intuition on which a computer cannot think or have understanding [4]. Ned Block states that Searle's argument depends for its force on intuitions that certain entities do not think. But this begs a question: Is this kind of intuition a correct one, as the progress of science changes our intuitions (the sun does not orbit anymore around the earth). Therefore, it seems to be impossible to settle these questions without employing a definition of the term 'understand' that can provide a test for judging the hypothesis is true or not.

AI researchers Herlbert Simon and Stuart Eisenstadt argue that various attributions of mentality—such as understanding—can be associated with programs by using "intentions" that determine extensions. Daniel Dennett claims the CRA is "clearly a fallacious and misleading argument..." For example, the technology of autonomous robotic cars has proven the CRA to be invalid argument.

3 ISSUES ON SYNTAX AND SEMANTICS

One of the key points of Searle's claims against the SAI is that syntax is not sufficient for semantics, therefore programs cannot produce minds [4].

1. Programs are purely formal (syntactic).
2. Human minds have mental contents (semantics).
3. Syntax by itself is neither constitutive of, nor sufficient for, semantic content.
4. Therefore, programs by themselves are not constitutive of nor sufficient for minds.

This argument relies on the linguistic distinction between syntax and semantics (syntactical properties and semantic properties). Searle points out that *"formal symbols by themselves can never be enough for mental contents, because the symbols, by definition, have no meaning (or interpretation, or semantics) except insofar as someone outside the system gives it to them."* This is an interesting aspect, because it begs questions: *What is the significant difference between a human learning a symbol and a meaning of the symbol and an AI application (for example a robot) learning a symbol and a meaning of the symbol? What is the significant difference between a human acting based on the meaning of the symbol correctly and an AI application acting based on the meaning of the symbol correctly?*

What does the first premise actually mean? It is not quite clear what the term *program* means? In computer science, a program is a specific set of operations for a computer to perform. As such, the program is not an entity that has any active or causal role in its environment; it is just a combination of bits in a computer memory or a print on a paper sheet. Thus, I agree that in this sense of the term *program* the first premise is true. But, this changes, if Searle means with the term *program* a running program, which is called *process* in computer science[4]. A process in a robot can have causal relationships with its environment, and there exist counter-factual worlds; thus the robot may have semantics[5]. But

4 We can assume that this is the case, as Searle mentions "instantiating a computer program" in the abstract of his article.

5 It is quite possible to build a robot that can be taught to order and bring me a pizza instead of a hamburger from a restaurant.

now, Searle argues that it is not the right kind of the causal relationship [9].

Let us have a following thought experiment: We have two exactly similar rooms, which have equipment to carry out the following training session. On the floors there are several tools, such as a screwdriver, a hammer, a saw, etc.. In one room there is a child to whom a man teaches the tools, and in another room there is a robot[6] to which another man teaches the tools. The teaching session goes in the following way: The teacher says to the child/robot **"Bring me hammer."**. As, at the first time, the child/robot does not know which tool is **hammer**, she/it picks up an arbitrary object which happens to be the **screwdriver**, and brings it to the teacher. The teacher says in both cases: "No, this is not **hammer**; this is **screwdriver**. The child/robot takes the **screwdriver** back to its place. Next the child/robot takes the **hammer** and brings it to the teacher. The teacher says in both cases: "Yes, this is **hammer**, thank you". Next the teacher says to the child/robot **"Bring me screwdriver."**. Now, the child/robot does know, which tool is the **screwdriver**, she/it brings the **screwdriver**. The teacher says in both cases: "Thank you".[7]

Now, according to Searle, the child (mind) has a mental content (semantics), but the robot does not have a mental content (semantics thus understanding). The robot only simulates[8] understanding of the meanings of the symbols, and there is not the right kind of the causal relationship[9] [9]. My intuition says that in both cases there is semantics involved. Searle's statement in context of the Brain Simulator Reply "*It would be reasonable to attribute understanding to an android system, but only as long as you don't know how it works. As soon as you know the truth—it is a computer manipulating symbols on the basis of syntax, not meaning–you would cease to attribute intentionality to it*." is interesting indicating that Searle is attributing to human some mystique properties (e.g. a hidden property of understanding, which is Searle's account of non-formal causal

6 The robot has required visual tools to recognize various objects, and limbs to move and catch objects.

7 This though scenario can be extended by adding a teach session discussing, what can be done with *hammer* and *screwdriver*.

8 It cannot be duplication, because we don't know yet how brain actually produces understanding and what are the limits of understanding (mind).

9 According to Searle the causal relationship should be bottom-up relationship and not input-output relationship.

power of the brain created by a biological entity.) which cannot be reproduced artificially using a digital computer.[10]

The third premise raises a question about what Searle actually means with the term *"syntax by itself"*. I argue that it is quite feasible that we can have semantics expressed with a program (e.g. using semantic web languages) in a similar way as a human mind expresses semantics with a natural language, and semantics expressed in this way gets its content in the execution of the program. Today's examples of a robot developing its own language defining terms about its surroundings and teaching the language to another robot [7], and of an application based on AI that wins the best human players in quiz game called Jeopardy prove that programs (a.k.a process) might be constitutive of and sufficient for a kind of minds[11].

Thus, I argue that the above argument is either valid (the first interpretation of the term *program*), but it says nothing about modern applications based on AI, or it is not valid (the second interpretation of the term *program*).

David Cole and Daniel Dennett, among others, support the idea that a computer running a program is not the same as *syntax alone*. A computer is an enormously complex electronic causal system. Actually Dennett argues that programming is precisely what could give something a mind — but only on organic, human brains. He also argues that Searle has apparently confused a claim about the underivability of semantics from syntax with a claim about the underivability of the consciousness of semantics from syntax [10]. Georges Rey and David Chalmers argues that a realization is not just a structural mapping, but involves causation, supporting counterfactuals. *"This point is missed so often: the syntactically specifiable objects over which computations are defined can and standardly do possess a semantic: it is just that the semantics is not involved in specification."* [4].

10 Of course, one could consider counterfactually than when human being knows in the future the truth about how a mind is created by a human brain, one could cease to attribute intentionality to it ;-).

11 Please, see http://www.ibm.com/smarterplanet/us/en/ibmwatson/

4 ISSUES ABOUT INTENTIONALITY AND CONSCIOUSNESS

A lot of discussion about the capability of a computer to be a mind deals with intentionality. Intentionality is the power of mind to be about, to represent, or to stand for things, properties and states of affairs. It is a pervasive feature of many different mental states, such as beliefs, hopes, judgements, intentions, and love. [11] But intentionality is not yet well-understood, and there are many open issues; for example, how mind does create intentionality. Searle has not specified an account of intentionality that is precise enough, and until Searle does it, it is an open question whether the SAI could produce intentionality, or whether it is beyond its scope.

According to Searle consciousness is the necessary condition of intentionality, and we can interpret the states of computer as having content, but the states themselves do not have original intentionality. It is consciousness that is lacking in digital computers. [4] On the other hand, Ray Kurzweil argues that AI systems can potentially have such mental properties as understanding, intelligence, consciousness and intentionality, and will exceed human abilities in these areas[12]. And Daniel Dennett is of the opinion that all intentionality is derived, and attributions of intentionality are instrumental and allow to predict behaviour. Fred Dretske sees intentionality as information–based, and a state of the world may carry information about other states in the world. This informational aboutness is a mind–independent features of states. Searle claims that the kind of intentionality that computers apparently display is not the kind which humans display [3]. The level of consciousness varies quite a lot from human to human, from animal to animal and exhibits itself in different ways, thus there is not an ubiquitous phenomena called consciousness. As long as the philosophy of the mind and neuroscience cannot answer properly the questions *"how does mind create consciousness?"*, *"how is intentionality established?*, and *"what is real understanding?"* the claims of the CRA about consciousness remain unsolved.

12 Please see the book Ray Kurzweil, The Singularity is Near: When Humans Transcend Biology

5 SOME OTHER THOUGHTS

Searle argues that the Chinese Room thought experiment proves it to be untrue that *the machine can literally be the said to understand the story and provide the answers to questions* [2]. This claim is based on two fundamental logical truth: First, syntax is not semantics, and secondly, simulation is not duplication [5]. However, Searle does not explicate well enough what he means with the phrase *"literally be said to understand"*. According to WordNet[13] the word *"understand"* is explicated in the context of language as follows: *"make sense of a language"* [12]. I agree that the man in the Chinese Room does not make sense of Chinese. But I argue that the robot in the thought experiment in section 3 have learnt some English and does make sense of what it has learnt. Now, what might *"literally"* mean? Searle discusses the right kind of the causal relationship to be a form of 'bottom–up' causality, where the specific neurobiological processes in the brain establish intentionality, understanding, etc. Searle seems to claim that this kind of the causal relationship is possible only for biological entities; all other kinds of causal relationships (e.g. implemented using digital computers) are either wrong kinds—input–output—or a simulation of the biological causal relationship. Does the robot duplicate or simulate the capability of understanding exercised by the child when showing to make sense of the English utterance *"Bring me screwdriver."* by collecting the right item from the floor and taking it to the teacher? This is a difficult issue. My intuition says that it is duplication, because the causal result of *"understanding"* is same in the real world.[14] But I am not quite confident of my intuition. Based on this discussion I am of the opinion that Searle seems to have built a sandbox in which he can argue and be right that the SAI is false. But is the CRA meaningful in the domain of modern AI, anymore?

I see that both the Turing test and the Chinese Room thought experiment are outdated from the viewpoint of AI. But as the argument to refute the computationalist and functionalist theories of mind the CRA has the role in the philosophy of mind. The term *Strong AI* as it is specified by Searle is no longer relevant in the

13 A lexical database for English published by Princeton University.

14 An example: simulation of heart in a virtual world versus artificial heart actually taking care of the blood circulation.

context of AI, and AI people are more discussing about the term *superintelligence*. Superintelligence refers to artificial entities that greatly outperform the best current human minds in most general cognitive domains. As Nick Bostrom in his book *Superintelligence: Paths, Dangers, Strategies* states that an artificial intelligence need not much resemble a human mind. Artificial intelligence will have very different cognitive architectures than biological intelligences [13]. Therefore, instead of the Turing test and the CRA we should consider whether an artificial entity can act as a full member of a society fulfilling adequately its responsibilities as the member of the society (equally to a human member of the society)? For example, can there be a robot that could autonomously operate a cancer patient as well as a human surgeon, or can there be an avatar that can provide students with a lecture series about philosophy of mind at a university, or can there be an autonomous, intelligent, honest software agent that acts a politician in a parliament?

My personal opinion can be summarized in the following claims: 1) Externalist attitude: states of a physical entity get their content through causal connections to the external reality they represent (Fred Dretske, Hilary Putnam, and Jerry Fodor). This is not limited only to human beings. 2) A computer might have propositional attitudes if it has the right causal connections to the world. (Jerry Fodor). 3) The syntactically specifiable objects over which computations are defined can and standardly do possess a semantic; it's just that the semantics is not involved in the specifications (Georges Rey) [14]. For example, in the context of Intelligent Software Agents we can have semantics involved using metadata about information. 4) For example, intentional state, just as belief, can be split into two distinct properties: conscious awareness of the belief and intentional state; thus, to allow attribution of intentionality to systems that can learn (Fred Dretske). Intentional state is the key idea, here. 5) Programming is precisely what could give something a mind (Daniel Dennett).

6 SUMMARY

John R. Searle has succeeded in raising a lot of discussion of the capabilities of artificial intelligence to create human-like minds.

He argues that his Chinese Room Argument proves that instantiating a program is never either constitutive or sufficient for minds. The many issues raised by the Chinese Room argument may not be settled until there is a consensus about the nature of meaning, its relation to syntax, and about the biological basis of consciousness. There continue to be significant disagreement about which processes create meaning, understanding, and consciousness.

I see that the main weaknesses of John R. Searle's arguments are as follows: 1) Terms, such as understand, intention, and consciousness are not unambiguously explicated, well enough. Are we arguing about same issues? 2) John R. Searle seems assume that human beings have some kind of higher level metaphysical entity that explains the 'superiority' features over artificial entities. Therefore, the terms understand, intention, and consciousness are not possible for any kind of digital computers. 3) The Chinese Room Argument is outdated, nowadays.

Today's examples of the results of modern artificial intelligence indicate that the actual issue is not whether a digital computer with the right program can be a mind, but does superintelligence greatly outperform the best current human minds in most general cognitive domains?

Anyway, anthropomorphism still seems to be a difficult issue, especially from humanistic point of view, as Sir Anthony Kenny pointed out in his first Georg Henrik von Wright lecture at the University of Helsinki [15].

REFERENCES

[1] Jaegwon Kim. *Philosophy of Mind*. Westview Press, 2010. ISBN 9780813344584.

[2] John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457, 1980.

[3] John Preston. *Views into the Chinese Room - New Essay on Searle and Artificial Intelligence*. Claredon Press - Oxford, first edition, 2002. ISBN 0-19-925277-7.

[4] David Cole. The chinese room argument. In Edward N. Zalta, editor, *The Stanford Encyclopedia of*

Philosophy. Metaphysics Research Lab, Stanford University, spring 2019 edition, June 2019. Available on https://plato.stanford.edu/archives/spring2019/entries/chinese-room/.

[5] John R. Searle. *Twenty-One Years in the Chinese Room*, chapter Twenty-One Years in the Chinese Room, page 51–69. Claredon Press - Oxford, 2002. ISBN 0-19-925277-7.

[6] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 2009. ISBN 0-13-604259-7.

[7] Luc Steels. Can robots be made creative enough to invent their own language? January 2012. Available on https://www.youtube.com/watch?v=n6876onk7sl (01.09.2014).

[8] Stevan Harnad. Alan turing and the "hard" and "easy" problem of cognition: Doing and feeling. *Turing100: Essays in Honour of Centenary Turing Year 2012*, 2012. Available on http://eprints.soton.ac.uk/340293/1/harnad-humaturingessay.pdf.

[9] John R. Searle. Consciousness & the brain: John searle at tedxcern. Youtube, May 2013. URL: https://www.youtube.com/ John Searle presentation at TEDxCERN.

[10] Daniel. Fast thinking. In *The Intentional Stance*. The MIT Press, Cambridge, Massachusetts, 1996. ISBN 0-262-54053-3.

[11] Pierre Jacob. Intentionality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2019 edition, March 2019. Available on https://plato.stanford.edu/archives/spr2019/entries/intentionality/.

[12] Electronic. Available on https://wordnet.princeton.edu.

[13] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. ISBN 978-0-19-967811-2.

[14] Georges Rey. Searle's misunderstanding of functionalism and strong ai. In John Preston and Mark Bishop, editors, *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, chapter Searle's Misunderstanding of Functionalism and Strong AI, page 201–225. Oxford University Press, 2002. ISBN 0-19-825057-6.

[15] Sir Anthony Kenny. Anthropomorphism vs humanism. Georg Henrik von Wright Lecture at University of Helsinki, June 2014. Available on https://www.helsinki.fi/en/unitube/video/c2be9093-1305-4341-8d6b-6bddb40cd0d2.

TIETOJENKÄSITTELYTIETEEN OSASTO
PL 68 (Pietari Kalmin katu 5)
00014 Helsingin yliopisto

DEPARTMENT OF COMPUTER SCIENCE
P.O. Box 68 (Pietari Kalmin katu 5)
FI-00014 University of Helsinki, FINLAND

JULKAISUSARJA **A**

SERIES OF PUBLICATIONS **A**

Reports are available on the e-thesis site of the University of Helsinki.

A-2015-1 L. Wang: Content, Topology and Cooperation in In-network Caching. 190 pp. (Ph.D. Thesis)

A-2015-2 T. Niinimäki: Approximation Strategies for Structure Learning in Bayesian Networks. 64+93 pp. (Ph.D. Thesis)

A-2015-3 D. Kempa: Efficient Construction of Fundamental Data Structures in Large-Scale Text Indexing. 68+88 pp. (Ph.D. Thesis)

A-2015-4 K. Zhao: Understanding Urban Human Mobility for Network Applications. 62+46 pp. (Ph.D. Thesis)

A-2015-5 A. Laaksonen: Algorithms for Melody Search and Transcription. 36+54 pp. (Ph.D. Thesis)

A-2015-6 Y. Ding: Collaborative Traffic Offloading for Mobile Systems. 223 pp. (Ph.D. Thesis)

A-2015-7 F. Fagerholm: Software Developer Experience: Case Studies in Lean-Agile and Open Source Environments. 118+68 pp. (Ph.D. Thesis)

A-2016-1 T. Ahonen: Cover Song Identification using Compression-based Distance Measures. 122+25 pp. (Ph.D. Thesis)

A-2016-2 O. Gross: World Associations as a Language Model for Generative and Creative Tasks. 60+10+54 pp. (Ph.D. Thesis)

A-2016-3 J. Määttä: Model Selection Methods for Linear Regression and Phylogenetic Reconstruction. 44+73 pp. (Ph.D. Thesis)

A-2016-4 J. Toivanen: Methods and Models in Linguistic and Musical Computational Creativity. 56+8+79 pp. (Ph.D. Thesis)

A-2016-5 K. Athukorala: Information Search as Adaptive Interaction. 122 pp. (Ph.D. Thesis)

A-2016-6 J.-K. Kangas: Combinatorial Algorithms with Applications in Learning Graphical Models. 66+90 pp. (Ph.D. Thesis)

A-2017-1 Y. Zou: On Model Selection for Bayesian Networks and Sparse Logistic Regression. 58+61 pp. (Ph.D. Thesis)

A-2017-2 Y.-T. Hsieh: Exploring Hand-Based Haptic Interfaces for Mobile Interaction Design. 79+120 pp. (Ph.D. Thesis)

A-2017-3 D. Valenzuela: Algorithms and Data Structures for Sequence Analysis in the Pan-Genomic Era. 74+78 pp. (Ph.D. Thesis)

A-2017-4 A. Hellas: Retention in Introductory Programming. 68+88 pp. (Ph.D. Thesis)

A-2017-5 M. Du: Natural Language Processing System for Business Intelligence. 78+72 pp. (Ph.D. Thesis)

A-2017-6 A. Kuosmanen: Third-Generation RNA-Sequencing Analysis: Graph Alignment and Transcript Assembly with Long Reads. 64+69 pp. (Ph.D. Thesis)

A-2018-1 M. Nelimarkka: Performative Hybrid Interaction: Understanding Planned Events across Collocated and Mediated Interaction Spheres. 64+82 pp. (Ph.D. Thesis)

A-2018-2 E. Peltonen: Crowdsensed Mobile Data Analytics. 100+91 pp. (Ph.D. Thesis)

A-2018-3 O. Barral: Implicit Interaction with Textual Information using Physiological Signals. 72+145 pp. (Ph.D. Thesis)

A-2018-4 I. Kosunen: Exploring the Dynamics of the Biocybernetic Loop in Physiological Computing. 91+161 pp. (Ph.D. Thesis)

A-2018-5 J. Berg: Solving Optimization Problems via Maximum Satisfiability: Encodings and Re-Encodings. 86+102 pp. (Ph.D. Thesis)

A-2018-6 J. Pyykkö: Online Personalization in Exploratory Search. 101+63 pp. (Ph.D. Thesis)

A-2018-7 L. Pivovarova: Classification and Clustering in Media Monitoring: from Knowledge Engineering to Deep Learning. 78+56 pp. (Ph.D. Thesis)

A-2019-1 K. Salo: Modular Audio Platform for Youth Engagement in a Museum Context. 97+78 pp. (Ph.D. Thesis)

A-2019-2 A. Koski: On the Provisioning of Mission Critical Information Systems based on Public Tenders. 96+79 pp. (Ph.D. Thesis)

A-2019-3 A. Kantosalo: Human-Computer Co-Creativity - Designing, Evaluating and Modelling Computational Collaborators for Poetry Writing. 74+86 pp. (Ph.D. Thesis)

A-2019-4 O. Karkulahti: Understanding Social Media through Large Volume Measurements. 116 pp. (Ph.D. Thesis)

A-2019-5 S. Yaman: Initiating the Transition towards Continuous Experimentation: Empirical Studies with Software Development Teams and Practitioners. 81+90 pp. (Ph.D. Thesis)

A-2019-6 N. Mohan: Edge Computing Platforms and Protocols. 87+69 pp. (Ph.D. Thesis)

A-2019-7 I. Järvinen: Congestion Control and Active Queue Management During Flow Startup. 87+48 pp. (Ph.D. Thesis)

A-2019-8 J. Leinonen: Keystroke Data in Programming Courses. 56+53 pp. (Ph.D. Thesis)

A-2019-9 T. Talvitie: Counting and Sampling Directed Acyclic Graphs for Learning Bayesian Networks. 70+54 pp. (Ph.D. Thesis)

A-2019-10 J. Toivonen: Modeling and Learning Monomeric and Dimeric Transcription Factor Binding Motifs. 61+109 pp. (Ph.D. Thesis)

A-2019-11 S. Hemminki: Advances in Motion Sensing on Mobile Devices. 113+89 pp. (Ph.D. Thesis)

A-2019-12 P. Saikko: Implicit Hitting Set Algorithms for Constraint Optimization. 70+54 pp. (Ph.D. Thesis)

A-2020-1 J. Leppä-aho: Methods for Learning Directed and Undirected Graphical Models. 50+84 pp. (Ph.D. Thesis)

A-2020-2 P. Zhou: Edge-Facilitated Mobile Computing and Communication. 137 pp. (Ph.D. Thesis)

A-2020-3 J. N. Alanko: Space-Efficient Algorithms for Strings and Prefix-Sortable Graphs. 67+82 pp. (Ph.D. Thesis)

A-2020-4 H. Mäenpää: Organizing and Managing Contributor Involvement in Hybrid Open Source Software Development Communities. 78+67 pp. (Ph.D. Thesis)